Coordinated Multicasting with Opportunistic User Selection in Multicell Wireless Systems

Y.-W. Peter Hong, Wei-Chiang Li, Tsung-Hui Chang, and Chia-Han Lee

Abstract

Physical layer multicasting with opportunistic user selection (OUS) is examined in this work for multicell multi-antenna wireless systems. In multicast applications, a common message is to be sent by the base stations to all users in a multicast group. By adopting a two-layer encoding scheme, a rate-adaptive channel code is applied in each fading block to enable successful decoding by a chosen subset of users (which varies over different blocks) and an application layer erasure code is employed across multiple blocks to ensure that every user is able to recover the message after decoding successfully in a sufficient number of blocks. The transmit signal and code-rate in each block determine opportunistically the subset of users that are able to successfully decode and can be chosen to maximize the long-term multicast efficiency. The employment of OUS not only helps avoid rate-limitations caused by the user with the worst channel, but also helps coordinate interference among different cells and multicast groups. In this work, efficient algorithms are proposed for the design of the transmit covariance matrices, the physical layer code-rates, and the target user subsets in each block. In the single group scenario, the system parameters are determined by maximizing the group-rate, which is defined as the physical layer code-rate times the fraction of users that can successfully decode in each block. In the multi-group scenario, the system parameters are determined by considering a group-rate balancing optimization problem (i.e., a max-min weighted group-rate problem), which is solved using a successive convex approximation (SCA) approach. To further reduce the feedback overhead, we also consider the case where only part of the users feed back their channel vectors in each block and propose a design based on the balancing of the expected group-rates. In addition to applying SCA, a sample average approximation technique is also introduced to handle the probabilistic terms that arise in this problem. The effectiveness of the proposed schemes is demonstrated through computer simulations.

Y.-W. P. Hong and W.-C. Li (emails: ywhong@ee.nthu.edu.tw and weichiangli@gmail.com) are with the Institute of Communications Engineering, National Tsing Hua University, Hsinchu, Taiwan. T.-H. Chang (email: tsunghui.chang@ieee.org) is with the Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan. C.-H. Lee (email: chiahan@citi.sinica.edu.tw) is with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan.

This work was supported in part by the Ministry of Science and Technology under grants 100-2628-E-007-025-MY3 and 102-2221-E-007-016-MY3.

I. Introduction

A. Motivation and Background

Multicasting has attracted much attention in recent years due to the increasing demand for mass content distribution, such as software and firmware updates, file downloads, and multimedia streaming. In these applications, common information is to be disseminated efficiently to all users in a multicast group. Due to its importance in mobile applications, such service is also being introduced into current and next-generation cellular standards, such as the global system for mobile communications (GSM), the worldwide interoperability for microwave access (WiMAX), the long term evolution (LTE) etc., in the form of the so-called multimedia broadcast/multicast service (MBMS) [1]–[3]. Different from wireline networks, multicasting in wireless systems enjoys the so-called wireless broadcast advantage (namely, the advantage that all users within the transmission range can receive) and, therefore, many research efforts have been devoted to the development of physical layer techniques that can fully exploit these advantages.

Most works in the literature on wireless physical layer multicasting consider the transmit signal design at the base station (BS) or BSs for efficient delivery of common information to all users in a multicast group. Due to fading, the channels experienced by users in the multicast group may vary drastically and, thus, the rate of the channel code must be low enough to ensure that all users in the group can successfully decode. When multiple antennas are available at the transmitter, beamforming and precoding techniques can be further employed to improve the effective channel quality of the worst user [4]–[6] and, thus, the multicast rate. With no restrictions on the rank of the signal covariance matrix, the optimal precoder (or the transmit covariance matrix) can be found using semi-definite programming (SDP) techniques [7]. However, to reduce decoding complexity at the receivers, many works, such as [4] and [5], considered instead the use of multicast beamforming (i.e., signals with rank-1 covariance matrices) and proposed approximate algorithms for the design of multicast beamformers, which is otherwise known to be NP-hard [4]. These works were extended to systems with multiple multicast groups in [8]–[12], where cochannel interference between different groups' messages was taken into consideration, and were also extended to multicell systems in [13], [14], where coordinated transmissions among BSs were employed. However, even with the above signal designs, the efficiency of the physical-layer multicast transmission is still fundamentally limited by the user with the worst channel

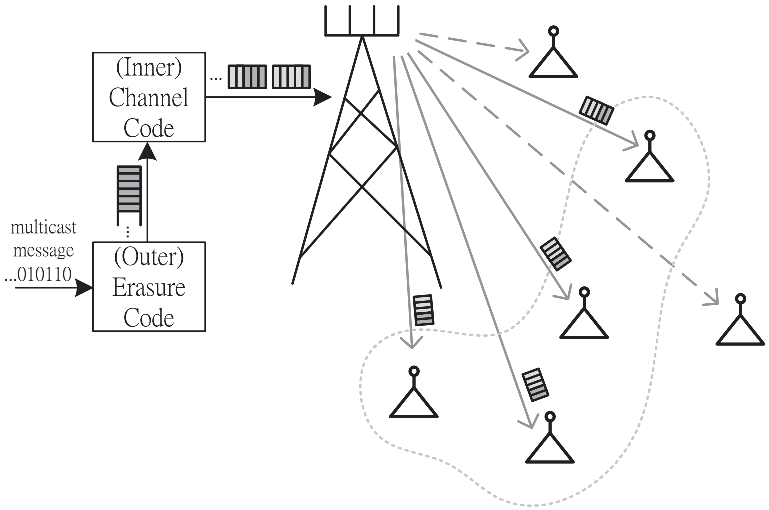

Fig. 1. Illustration of the opportunistic multicast scheduling (OMS) scheme.

and the rate required for all users to decode may eventually go to zero as the number of users in the group increases.

B. Related Works on Opportunistic Multicasting

For delay-tolerant applications, such as file downloads and software or firmware updates, the limitations caused by the worst user in the network can be overcome by dividing the transmission into multiple blocks and by scheduling only a subset of users to receive in each block. This technique was referred to as *opportunistic multicast scheduling (OMS)* in [15]–[18]. In particular, the OMS scheme requires a two-layer encoding scheme, as illustrated in Fig. 1, which consists of a physical layer channel code (hereby referred to as the inner code) whose rate is adapted to the chosen user subset in each block and an application layer erasure code (referred to as the outer code) that is performed across multiple blocks. By selecting only a subset of users to serve in each block, the rate of the inner code would be less restrictive since successful decoding needs to be guaranteed for fewer users in each block. The overall *group-rate*, defined as the code-rate multiplied by the fraction of users in the group that successfully decode, can thus be effectively increased. However, since a user may not be able to successfully decode in every block, the channel that it experiences across multiple blocks is effectively an erasure channel, and an outer erasure code should be employed to ensure that all users in the multicast group can eventually recover the message after decoding over a sufficient number of blocks. The OMS scheme can achieve significant performance gains, but comes at the expense of delay. That is, these gains result from using time as an additional resource or degree of freedom as compared to conventional schemes that require successful and instantaneous decoding in every block.

In practice, the outer code can be implemented using, e.g., LT, Raptor, and Fountain codes [19], [20]. With OMS, the group-rate is known to converge to a non-zero constant as the number of users inceases [15]–[17]. This is in contrast to cases without OMS where the group-rate is known to diminish to zero. Most works in the literature on OMS focused on the single-cell single-antenna scenario, as in [15]–[18]. Our work focuses instead on the multi-antenna scenario, where the problem is considerably more difficult due to the dependence between the user selection and the transmit signal design. The multi-antenna scenario was also examined more recently in [21]–[23], but only for single-cell scenarios. In particular, in [21], the joint precoding and user selection was performed using a heuristic semi-orthogonal vector selection algorithm. Their proposed scheme has low complexity but is suitable only for single-cell scenarios with sum power constraints. In [22], [23], the transmit signal design was restricted to beamforming and the user selection was performed using a heuristic subset search algorithm. The algorithm can be extended to multicell scenarios, but is subject to high computational complexity.

C. *Main Contributions*

The main contribution of this work is the development of efficient algorithms for the joint design of the transmit covariance matrix and the opportunistic user selection policy in multicell networks with multiple multicast groups. Here, group-rate (which is defined as the physical layer code-rate in a certain block multiplied by the fraction of users that can successfully decode) is utilized as the optimizing criterion. This is different from most works in the literature on physical layer multicasting, e.g., [4], [8], which typically do not consider user selection and, thus, can utilize the signal-to-interference-plus-noise ratio (SINR) as the optimizing criterion. In fact, the SINR criterion is not applicable when user selection is considered since it does not reflect the effect of the number of users served in each block. For example, if the system is optimized by maximizing the worst SINR among all selected users, then only one user (i.e., the user with the best effective channel) in each group would be selected. However, finding the optimal transmit covariance matrix and the opportunistic user selection (OUS) policy to maximize the group-rate may be difficult and is, in fact, claimed to be NP-hard in [22]. Therefore, we propose in this work an approximate solution based on the introduction and relaxation of a set of binary user selection variables. This technique is similar to that previously adopted in [24] and [25] for admission control problems.

In this work, we consider both single-group and multi-group multicasting scenarios. We show that OUS can be especially effective in the latter case, where it not only can help avoid limitations by the worst user but also can help coordinate interference among different cells and multicast groups. In the single group scenario, the optimization of the transmit covariance matrix and the user subset selection is performed first by relaxing the integer constraints corresponding to the user selection variables and then by performing a sequential deflation technique, where users are eliminated one-by-one from an initial set of all users to yield candidate user subsets. In the multi-group scenario, the design parameters are determined based on the group-rate balancing criterion (also referred to as the max-min weighted group-rate criterion) and the optimization problem is solved using a similar relaxation technique along with a successive convex approximation (SCA) approach. In the above, the channel state information (CSI) of all users is first assumed to be available at the transmitter, which may be costly in practice. To reduce the transmission overhead, we also consider the case where only a subset of users feeds back their CSI in each block and propose a design based on the balancing of the expected group-rates. In addition to the SCA approach mentioned above, a sample average approximation (SAA) technique [26], [27] is further introduced to handle the probabilistic terms that may arise. The effectiveness of the proposed schemes, compared to [21]–[23], is demonstrated through computer simulations.

In multicell systems, three cases are often considered based on different levels of BS cooperation [28]–[30], namely, full BS cooperation, interference coordination, and no BS cooperation. *Full BS cooperation* refers to the case where all BSs have knowledge of the information intended to all multicast groups and transmit cooperatively as a networked multiple-input multiple-output (MIMO) system. *Interference coordination* refers to the case where each BS serves only one multicast group and only has knowledge of the message intended for that group. In this case, cooperative transmission of common data is not possible, but the transmit signals can be designed to reduce cochannel interference among different multicast groups. *No BS cooperation* refers to the case where no CSI from users in other cells is available and thus no cooperation is adopted. Our system model is general enough to include all the cases mentioned above. Moreover, we

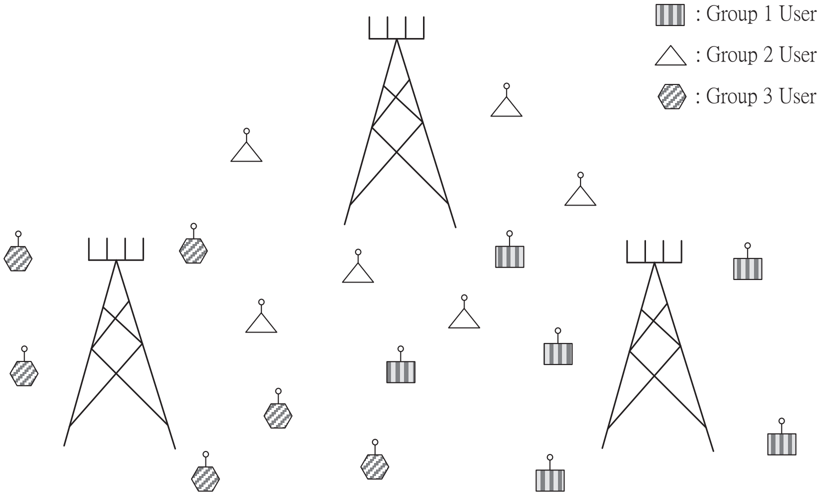

Fig. 2. Illustration of a multicell network with 3 multicast groups.

would also like to mention that, in this work, we do not restrict ourselves to rank-1 transmissions, as done in [4], [5], [8]–[10], [22], [23], in order to maintain optimality [7] and also to avoid distracting the readers, since a rank-1 constraint on the transmit covariance matrix is already enough to make the problem intractable [4], even without considering user selection. However, if rank-1 transmissions are desired, our solution can be used as the semi-definite relaxation (SDR) of the optimal beamformer design and the desired beamforming vectors can be extracted from our solutions using rank-1 approximation techniques, e.g., the Gaussian randomization procedure, as discussed in [31].

The remainder of the paper is organized as follows. In Section II, we introduce the general multicell multicast scenario and the proposed group-rate balancing problem. In Sections III and IV, the joint design of the transmit covariance matrix and the user selection policy is examined for the single-group and the multi-group scenarios, respectively. In Section V, the problem is extended to the case where only a subset of users feeds back their CSI in each block. Finally, computer simulations are provided in Section VI and the paper is concluded in Section VII.

II. System Model and Problem Formulation

Let us consider a downlink multicell network with B BSs, each equipped with M antennas, and K single-antenna users. An illustration of the system under consideration is given in Fig. 2. The set of users $\mathcal{K}$, with cardinality $|\mathcal{K}| = K$, is divided into G multicast groups, denoted by $\mathcal{K}_1$, ..., $\mathcal{K}_G$, where users in the same group are interested in receiving the same multicast message. Moreover, let $\mathcal{B}_g \subseteq \{1, \ldots, B\}$ be the set of BSs that are serving group $\mathcal{K}_g$ and let $\mathcal{G}_b \subseteq \{1, \ldots, G\}$ be the set of multicast groups served by BS b. Notice that the case of full BS cooperation can be considered by setting $\mathcal{B}_g = \{1, \ldots, B\}$, for all g, and $\mathcal{G}_b = \{1, \ldots, G\}$, for all b, whereas the case of interference coordination, where each BS has information intended for only one multicast group, can be considered by setting $B = G$, $\mathcal{B}_g = \{g\}$ and $\mathcal{G}_b = \{b\}$, for all g and b. The case of no cooperation can also be considered by setting $B = G = 1$, $\mathcal{B}_1 = \{1\}$ and $\mathcal{G}_1 = \{1\}$, where the BS focuses on sending information to the intended users without knowledge of other BSs' channels or transmitted signals (i.e., while neglecting the presence of other BSs).

Here, we consider a delay-tolerant application where multicast messages can be encoded over multiple coherence intervals (hereafter referred to as *blocks*). Let $\mathbf{s}_{b,g}[n]$ be the signal transmitted by BS $b \in \mathcal{B}_g$ to users in group g in block n. In this case, the received signal at user $k \in \mathcal{K}_g$ can be written as

$$y_k[n] = \sum_{b \in \mathcal{B}_g} \mathbf{h}_{b,k}[n]^H \mathbf{s}_{b,g}[n] + \sum_{b \in \mathcal{B}_g} \mathbf{h}_{b,k}[n]^H \sum_{\substack{\ell \in \mathcal{G}_b \\ \ell \neq g}} \mathbf{s}_{b,\ell}[n] + \sum_{b' \notin \mathcal{B}_g} \mathbf{h}_{b',k}[n]^H \sum_{\ell \in \mathcal{G}_{b'}} \mathbf{s}_{b',\ell}[n] + n_k[n], \quad (1)$$

where $\mathbf{h}_{b,k}[n]$ is the $M \times 1$ complex channel vector between BS b and user k in block n and $n_k[n]$ is the complex additive white Gaussian noise (AWGN) at user k with zero mean and unit variance, i.e., $n_k \sim \mathcal{CN}(0, 1)$. The entries of $\mathbf{h}_{b,k}$ are assumed to be independent complex Gaussian random variables with variances that may be different for different b and k, due to path loss or other large-scale fading effects. The first term in (1) represents the sum of signals intended for user k, the second term represents the interference caused by the signals transmitted by user k's serving BSs to users in other multicast groups, and the third term represents the interference caused by non-serving BSs.

By defining $\mathbf{h}_k[n] = [\mathbf{h}_{1,k}[n]^H, \ldots, \mathbf{h}_{B,k}[n]^H]^H$ and $\mathbf{s}_g = [\mathbf{s}_{1,g}[n]^H, \ldots, \mathbf{s}_{B,g}[n]^H]^H$, where $\mathbf{s}_{b,g} = \mathbf{0}$ for $b \notin \mathcal{B}_g$, the received signal at user k can be rewritten as

$$y_k[n] = \mathbf{h}_k[n]^H \mathbf{s}_g[n] + \mathbf{h}_k[n]^H \sum_{\ell \neq g} \mathbf{s}_\ell[n] + n_k[n]. \quad (2)$$

Therefore, the SINR at user $k \in \mathcal{G}_g$ in block n is given by

$$\mathrm{SINR}_k[n] = \frac{\mathbf{E}[|\mathbf{h}_k[n]^H \mathbf{s}_g[n]|^2]}{\sum_{\ell \neq g} \mathbf{E}[|\mathbf{h}_k[n]^H \mathbf{s}_\ell[n]|^2] + 1} = \frac{\mathrm{tr}(\mathbf{Q}_g[n] \mathbf{h}_k[n] \mathbf{h}_k[n]^H)}{\sum_{\ell \neq g} \mathrm{tr}(\mathbf{Q}_\ell[n] \mathbf{h}_k[n] \mathbf{h}_k[n]^H) + 1} \quad (3)$$

where $\mathbf{Q}_g[n] = \mathbf{E}[\mathbf{s}_g[n]\mathbf{s}_g[n]^H]$. Note that $\mathbf{Q}_g[n]$ is an $MB \times MB$ matrix whose (b, b')-th block entry of dimension $M \times M$ is $\{\mathbf{Q}_g[n]\}_{b,b'} \triangleq \mathbf{E}[\mathbf{s}_{b,g}[n]\mathbf{s}_{b',g}[n]^H]$. This block entry is zero if either b or b' does not belong to $\mathcal{B}_g$. The signals transmitted by BS b must satisfy the per BS power constraint

$$\sum_{g\in\mathcal{G}_b} \mathbf{E}[\|\mathbf{s}_{b,g}[n]\|^2] = \sum_{g\in\mathcal{G}_b} \mathrm{tr}(\{\mathbf{Q}_g[n]\}_{b,b}) \leq P_b. \tag{4}$$

By employing OUS, only a subset of users in each group, i.e., $\mathcal{A}_g[n] \subseteq \mathcal{K}_g$, for group g, is required to decode successfully in block n. In this case, the rate of the physical layer encoding (i.e., the inner code) of group g, i.e., $R_g[n]$, only needs to be low enough to ensure successful decoding by users in $\mathcal{A}_g[n]$. More specifically, by assuming that the duration of each block is large enough to invoke Shannon's random coding argument [32], the rate $R_g[n]$ of the message intended for the subset of users $\mathcal{A}_g[n]$ in block n must be chosen such that

$$R_g[n] \leq \log\left(1 + \mathrm{SINR}_k[n]\right) = \log\left(1 + \frac{\mathrm{tr}(\mathbf{Q}_g[n]\mathbf{h}_k[n]\mathbf{h}_k[n]^H)}{\sum_{\ell\neq g}\mathrm{tr}(\mathbf{Q}_\ell[n]\mathbf{h}_k[n]\mathbf{h}_k[n]^H) + 1}\right) \ (\mathrm{bps/Hz}) \tag{5}$$

for all $k \in \mathcal{A}_g[n]$.

Since the codeword in each block is decoded only by a subset of users, each user effectively experiences an erasure channel across multiple blocks and, thus, an outer (erasure) code, such as LT, Raptor, and Fountain codes [19], [20], is needed to ensure eventual recovery of the original multicast data at all users. By assuming that an ideal erasure code is employed so that perfect erasure correction is performed, the average rate of user $k \in \mathcal{K}_g$ over T blocks can be written as

$$\frac{1}{T}\sum_{n=1}^{T} R_g[n]\mathbf{1}_{\{k\in\mathcal{A}_g[n]\}}, \tag{6}$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function, and the achievable multicast rate of group g is thus given by

$$\bar{R}_{g,\min} = \min_{k\in\mathcal{K}_g} \frac{1}{T}\sum_{n=1}^{T} R_g[n]\mathbf{1}_{\{k\in\mathcal{A}_g[n]\}}. \tag{7}$$

Conventional physical layer multicast problems often use $\bar{R}_{g,\min}$ as the maximization criterion for system design. However, maximizing $\bar{R}_{g,\min}$ would require non-causal knowledge of the channel realizations in the future T fading blocks, which is not attainable in practice. In order to implement the user selection and transmit signal design in real-time, we propose to maximize the (average) group-rate defined as

$$\bar{R}_g = \frac{1}{|\mathcal{K}_g|}\sum_{k\in\mathcal{K}_g}\frac{1}{T}\sum_{n=1}^{T}R_g[n]\mathbf{1}_{\{k\in\mathcal{A}_g[n]\}} = \frac{1}{T}\sum_{n=1}^{T}\bar{R}_g[n], \tag{8}$$

for group g, where

$$\bar{R}_g[n] = \frac{1}{|\mathcal{K}_g|}R_g[n]\sum_{k\in\mathcal{K}_g}\mathbf{1}_{\{k\in\mathcal{A}_g[n]\}} \tag{9}$$

is the instantaneous group-rate in block n. The average group-rate provides a measure of the average speed for which the users are able to acquire the multicast message. One can see that the maximization of the average group-rate can be achieved by maximizing the instantaneous group-rate block-by-block, and hence can be implemented in real-time. However, it should be noted that maximizing the average group-rate $\bar{R}_g$ may not be equivalent to maximizing the actual multicast rate $\bar{R}_{g,\min}$ in the general case. Therefore, $\bar{R}_g$ should only be viewed as an *auxiliary* maximization criterion that is used to allow for real-time optimization of the system parameters. However, it was shown in [22] (and observed similarly for single antenna scenarios in [15] and [18]) that, when the channel vectors are independent and identically distributed (i.i.d.) across users and over time, the multicast rate $\bar{R}_{g,\min}$ is asymptotically equal to the average group-rate $\bar{R}_g$ and, thus, is maximized asymptotically by maximizing the instantaneous group-rate $\bar{R}_g[n]$ in each block. The i.i.d. assumption is reasonable in many cases and is often considered when prior knowledge of the channel distributions or the users' locations are unavailable. When the channel vectors are non-i.i.d., normalization of the channel coefficients can be performed to ensure long-term fairness, as to be discussed in Sections III-B and IV-B.

Maximizing the instantaneous group-rate for block n involves optimizing over the transmit covariance matrices $\{\mathbf{Q}_g[n]\}_{g=1}^{G}$, the user subsets $\{\mathcal{A}_g[n]\}_{g=1}^{G}$, and the physical layer code-rates $\{R_g[n]\}_{g=1}^{G}$ subject to the power and rate constraints given in (4) and (5). This problem is examined for both the single-group and the multi-group multicasting scenarios. Since the optimization is performed separately in each block, we shall omit the block index n in the remainder of this work.

In the single-group multicasting scenario (i.e., the case where $G = 1$), the optimal system parameters can be determined by solving the following instantaneous group-rate maximization (GRM) problem.

Group-Rate Maximization (GRM) Problem:

$$\begin{aligned} \text{maximize} \quad & \frac{1}{|\mathcal{K}|} R \sum_{k\in\mathcal{K}} \mathbf{1}_{\{k\in\mathcal{A}\}} && \text{(10a)}\\ \text{subject to} \quad & \log_2[1+\text{tr}(\mathbf{Q}\mathbf{h}_k\mathbf{h}_k^H)] \geq R, \quad \text{for } k\in\mathcal{A}, && \text{(10b)}\\ & \text{tr}(\{\mathbf{Q}\}_{b,b}) \leq P_b,\ \forall b, \quad \mathbf{Q}\succeq \mathbf{0}. && \text{(10c)}\\ \text{variables:} \quad & \mathbf{Q}, \mathcal{A}, R. \end{aligned}$$

Notice that no cochannel interference exists in this case and that the group index g is omitted since only a single multicast group is considered. This problem is claimed to be NP-hard in [22] and no known solutions are available to solve this problem efficiently. An approximate solution will be studied in Section III based on the introduction of a set of binary user selection variables and a convex relaxation of the problem.

In the multigroup multicasting scenario (i.e., the case where $G > 1$), we consider the following group-rate balancing (GRB) optimization problem where the system parameters are jointly determined to maximize the worst weighted group-rate among all multicast groups.

Group-Rate Balancing (GRB) Problem:

$$\begin{aligned} \text{maximize} \quad & \min_{g\in\{1,\dots,G\}} \frac{1}{\tau_g}\frac{1}{|\mathcal{K}_g|} R_g \sum_{k\in\mathcal{K}_g} \mathbf{1}_{\{k\in\mathcal{A}_g\}} && \text{(11a)}\\ \text{subject to} \quad & \log_2\left(1+\frac{\text{tr}(\mathbf{Q}_g\mathbf{h}_k\mathbf{h}_k^H)}{\sum_{\ell\neq g}\text{tr}(\mathbf{Q}_\ell\mathbf{h}_k\mathbf{h}_k^H)+1}\right) \geq R_g, \forall k\in\mathcal{A}_g,\ \forall g, && \text{(11b)}\\ & \sum_{g=1}^{G}\text{tr}(\{\mathbf{Q}_g\}_{b,b}) \leq P_b,\ \forall b, \quad \mathbf{Q}_g\succeq\mathbf{0},\ \forall g, && \text{(11c)}\\ & \{\mathbf{Q}_g\}_{b,b'} = \mathbf{0}_{M\times M},\ \text{for } b\notin\mathcal{B}_g \text{ or } b'\notin\mathcal{B}_g, && \text{(11d)}\\ \text{variables:} \quad & \{\mathbf{Q}_g\}_{g=1}^G, \{\mathcal{A}_g\}_{g=1}^G, \{R_g\}_{g=1}^G, \end{aligned}$$

where τ_g is the parameter that specifies the priority of group $\mathcal{K}_g$. Solving the above problem guarantees a common level of quality-of-service (QoS) for all multicast groups. In particular, if the resulting objective value is α, then the group-rate of group g will be at least $\tau_g\alpha$, for all g. It is worthwhile to note that the GRB problem is analogous to the SINR-balancing problem often considered for the design of multiuser downlink beamforming [33] or multigroup multicast beamforming [8] schemes. While the SINR criterion may be suitable for conventional multicast beamforming designs, where all users are required to decode successfully in each fading block, it is not suitable for systems employing OUS. This is because, when considering the SINR-balancing (or, equivalently, the max-min SINR) criterion, the optimization will result in a trivial user selection policy where only the user with the best channel in each group is chosen. Even though the physical layer code-rate can be chosen to be the highest in this case, the group-rate is limited by the fact that only one user is able to decode. The GRB formulation is interesting in the sense that it allows us to take into consideration the effect of user selection as well as to incorporate the impact of QoS requirements, i.e., τ_g, into the problem. This problem will be examined further in Section IV and an approximate solution will be proposed based on a relaxation similar to that in the single-group scenario.

REMARK 1: *Most works in the literature on physical layer multicasting (e.g., [4]–[6], [8]–[10], [13], [14]) assume that all users must decode successfully in each block. In this case, the outer erasure code is not needed and the instantaneous group-rate will be equal to the physical layer code-rate in each block. The group-rate in block n must then satisfy*

$$\bar{R}_g[n] = R_g[n] \leq \log\left(1 + \mathrm{SINR}_k[n]\right), \quad \forall k \in \mathcal{K}_g.$$

The conventional approach can be viewed as the case where we choose $\mathcal{A}_g[n] = \mathcal{K}_g$, for all g and n. The proposed OUS scheme instead optimizes over $\mathcal{A}_g[n]$ and, thus, the achievable group-rate can be no less than that in conventional systems.

III. GROUP-RATE MAXIMIZATION FOR SINGLE-GROUP MULTICASTING

In this section, we examine the GRM problem in (10) for the single-group multicasting scenario. Note that the joint design of all system parameters in this problem is in general NP-hard [22] due to the combinatorial nature of the user subset selection. However, as similarly observed in [22], the problem reduces to a standard SDP problem when the user subset is given. In particular, for a given user subset $\mathcal{A}$, the SDP problem can be formulated as follows:

$$\begin{aligned} \max \quad & R \cdot |\mathcal{A}| && \text{(12a)} \\ \text{subject to} \quad & \log_2[1 + \mathrm{tr}(\mathbf{Q}\mathbf{h}_k\mathbf{h}_k^H)] \geq R, \quad \text{for } k \in \mathcal{A}, && \text{(12b)} \\ & \mathrm{tr}(\{\mathbf{Q}\}_{b,b}) \leq P_b, \ \forall b, \quad \mathbf{Q} \succeq \mathbf{0}, && \text{(12c)} \\ \text{variables:} \quad & \mathbf{Q}, R. \end{aligned}$$

Due to this observation, the authors in [22] proposed a heuristic subset search algorithm where all users are considered initially and one user is excluded from the set in each iteration until no further gain in the objective value is obtained. The user removed in each iteration is the user whose removal from the set results in the maximum increase in the objective value. Let $g(\mathcal{A})$ be the optimal objective value of (12) for a given user subset $\mathcal{A}$. Then, the subset search algorithm can be summarized as follows.

Algorithm 1 Subset Search Algorithm [22]

```
1) Set $\mathcal{A} = \{1, \ldots, K\}$.
2) Define $\mathcal{A}_{-k} \triangleq \mathcal{A} - \{k\}$, for all $k \in \mathcal{A}$, and let $k^* = \arg\max_k g(\mathcal{A}_{-k})$.
3) If $g(\mathcal{A}) > g(\mathcal{A}_{-k^*})$, then stop and take $\mathcal{A}$ as the solution; else, set $\mathcal{A} \leftarrow \mathcal{A} - \{k^*\}$ and go to Step 2.
```

Notice from Algorithm 1 that, in evaluating k^* in each iteration, the SDP problem in (12) needs to be solved for every possible choice of $\mathcal{A}_{-k}$. This requires one to solve the above SDP problem in the order of $O(K^2)$ times, which can be inefficient as K increases.

A. Reformulation and Relaxation of the GRM Problem using User Selection Variables

To solve the GRM problem in (10) more efficiently, let us first introduce a set of binary user selection variables $\{s_k, \forall k \in \mathcal{K}\}$, where $s_k = 1$ if user k is selected (i.e., if $k \in \mathcal{A}$) and $s_k = 0$, otherwise. By choosing δ to be sufficiently small, the GRM problem can be equivalently formulated as

$$\begin{aligned}
\max \quad & \frac{1}{|\mathcal{K}|} R \sum_{k \in \mathcal{K}} s_k && \text{(13a)}\\
\text{subject to} \quad & \log_2[1 + \text{tr}(\mathbf{Q}\mathbf{h}_k\mathbf{h}_k^H)] + \delta^{-1}(1 - s_k) \geq R, \ \forall k \in \mathcal{K}, && \text{(13b)}\\
& \text{tr}(\{\mathbf{Q}\}_{b,b}) \leq P_b, \ \forall b, \quad \mathbf{Q} \succeq \mathbf{0}, && \text{(13c)}\\
& s_k \in \{0, 1\}, \ \forall k, && \text{(13d)}\\
\text{variables:} \quad & \mathbf{Q}, R, \{s_k\}_{k \in \mathcal{K}}.
\end{aligned}$$

The equivalence of the two problems is stated in the following lemma. The proof is given in Appendix A.

LEMMA 1: *For* $\delta \leq \left[\max_k \log_2(1+\sum_{b=1}^{B} P_b \|\mathbf{h}_{b,k}\|^2)\right]^{-1}$*,* $(\mathbf{Q}^*, R^*, \{s_k^*\}_{k\in\mathcal{K}})$ *is an optimal solution of the problem in* (13) *if and only if* $(\mathbf{Q}^*, R^*, \mathcal{A}^*)$*, where* $\mathcal{A}^* \triangleq \{k \in \mathcal{K} : s_k^* = 1\}$*, is an optimal solution of the GRM problem in* (10)*.*

Notice that the problem in (13) is non-convex due to the integer constraints on s_k's in (13d). To obtain an efficient solution, we consider a relaxation of the problem where the integer constraints on s_k's are replaced with the linear constraints $0 \leq s_k \leq 1$, for all k. By taking the logarithm of the objective and by omitting the irrelevant variables, the relaxed problem can be written as

$$\begin{aligned} \max \quad & \ln R + \ln \sum_{k\in\mathcal{K}} s_k && \text{(14a)} \\ \text{subject to} \quad & \log_2[1+\text{tr}(\mathbf{Q}\mathbf{h}_k\mathbf{h}_k^H)] + \delta^{-1}(1-s_k) \geq R,\ \forall k \in \mathcal{K}, && \text{(14b)} \\ & \text{tr}(\{\mathbf{Q}\}_{b,b}) \leq P_b,\ \forall b, \quad \mathbf{Q} \succeq \mathbf{0}, && \text{(14c)} \\ & 0 \leq s_k \leq 1, \forall k, && \text{(14d)} \\ \text{variables:} \quad & \mathbf{Q}, R, \{s_k\}_{k\in\mathcal{K}}. \end{aligned}$$

This problem is convex and can be solved efficiently using general purpose solvers such as CVX [34]. By relaxing the integer constraints on the user selection variables, the term $\delta^{-1}(1-s_k)$ can take on any value between 0 and δ^{-1}, and can be viewed as a measure of rate violation, i.e., the difference between the code-rate R and the achievable rate of user k. Similar approaches have also been used in studies of admission control problems in [24] and [25]. Notice that the solution of s_k in the relaxed problem may not take on the value 0 or 1. To convert the solution of the relaxed problem to a feasible solution of (13), we propose a sequential deflation technique, where the D users with the D smallest values of s_k are removed in each iteration, after which the values of $\mathbf{Q}$, R, and s_k, $\forall k$, are updated by solving (14) again. The value of D can be chosen as 1 in most cases but can be chosen to be greater than one for complexity reduction. Among the $\lceil K/D \rceil$ possible subsets obtained from this procedure, where $\lceil a \rceil$ is the smallest integer not less than a, the user subset that yields the largest group-rate is chosen. The sequential deflation algorithm is summarized in Algorithm 2.

Notice that the sequential deflation algorithm proposed above is different from that proposed in [25] where the algorithm is to terminate whenever the removal of a user no longer results in an increase in group-rate. However, when applying their approach to our problem, the algorithm

Algorithm 2 Opportunistic User Selection by Sequential Deflation

```
(i) Initialize by setting $\mathcal{A} \leftarrow \mathcal{K}$, $\mathcal{A}^* \leftarrow \emptyset$, and $\alpha^* \leftarrow 0$.
(ii) Solve (12) for given user subset $\mathcal{A}$ and let $\tilde{\alpha}$ be the resulting objective value.
(iii) If $\tilde{\alpha} > \alpha^*$, then set $\alpha^* \leftarrow \tilde{\alpha}$ and $\mathcal{A}^* \leftarrow \mathcal{A}$.
(iv) Solve the relaxed problem (14) for $\mathcal{K} = \mathcal{A}$ to yield the values of $s_k$ for all $k \in \mathcal{A}$.
(v) Set $\mathcal{A} \leftarrow \mathcal{A} - \mathcal{S}_{\min}$, where $\mathcal{S}_{\min}$ is the set of users associated with the $D$ smallest values
    of $s_k$ among users in $\mathcal{A}$.
(vi) Repeat steps (ii)-(v) until $\mathcal{A} = \emptyset$. Then, take $\mathcal{A}^*$ as the desired user subset.
```

often terminates early in the process at a local optimum that is close to choosing all users as the serving subset. Moreover, it is worthwhile to remark that, in Algorithm 2, two convex optimization problems are solved in each iteration (i.e., one for computing the relaxed problem (14) and one for solving (12) for given $\mathcal{A}$) and the number of iterations is equal to the number of users K in the worst case, when $D = 1$. This is a significant improvement over the $O(K^2)$ worst-case complexity required in the subset search method summarized in Algorithm 1 [22].

B. *Fairness of the GRM Problem in the Non-I.I.D. Case based on Channel Normalization*

Notice that maximizing the instantaneous group-rate, as done in the previous subsection, equivalently maximizes the average group-rate in (8). When the channel vectors (i.e., the short-term fading coefficients) are i.i.d. across users, all users will have equal probability of being selected in each block and, thus, the average group-rate would eventually be the rate achieved by all users as $T \rightarrow \infty$ [22]. However, when the channel vectors are non-identically distributed, users with better average channels (namely, users closer to the BSs) may have a higher probability of being selected under our proposed scheme. In this case, an issue of fairness may arise, similar to that observed in [15] and [18] for the single-antenna case. In this section, a heuristic channel normalization technique is proposed to address the aforementioned fairness issue.

The key idea is to normalize the channel vectors of the users by their long term statistics and perform the proposed OUS based on the normalized channel vectors (which will then be i.i.d.). More specifically, suppose that the channel vector $\mathbf{h}_{b,k}[n]$ between BS b and user k in block n has entries that are i.i.d. $\mathcal{CN}(0, d_{b,k}^{-\alpha})$, where $d_{b,k}$ is the distance between BS b and user k and α

is the path loss exponent. Moreover, let $\bar{d}_b = \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}} d_{b,k}$ be the average distance of all users to BS b. Let us define the normalized channel vector between BS b and user k in block n as

$$\tilde{\mathbf{h}}_{b,k}[n] = \mathbf{h}_{b,k}[n]\sqrt{\frac{\bar{d}_b^{-\alpha}}{d_{b,k}^{-\alpha}}}, \tag{15}$$

which has entries that are i.i.d. $\mathcal{CN}(0, \bar{d}_b^{-\alpha})$. The normalized channel vectors are then utilized to compute the optimal user subset, denoted by $\tilde{\mathcal{A}}^*$, using the algorithm proposed in the previous subsection. The optimal transmit covariance matrix can then be computed by solving (12) for given $\tilde{\mathcal{A}}^*$ using the original channel vectors $\mathbf{h}_{b,k}$, $\forall b, k$.

The normalized channel vectors are scaled by their average distance so that the distance between the BSs and the users are taken into account in the signal-to-noise ratio (SNR) at the receiver. After normalization, the channel vectors experienced by all users will become i.i.d. and, by performing OUS based on the normalized channel vectors, all users will have equal probability of being selected in each block, ensuring fairness among users. The users that are opportunistically selected in each block will likely be users whose instantaneous channel gains are larger than their respective averages, exploiting the advantages of both temporal and multiuser diversity. It is also interesting to note that, even though the users in the subset $\tilde{\mathcal{A}}^*$ are chosen as target users and the transmit covariance matrix $\mathbf{Q}$ is chosen to maximize the rate of users in $\tilde{\mathcal{A}}^*$, it is possible that users outside of the target subset $\tilde{\mathcal{A}}^*$ may also be served due to their large average channel gains (i.e., their close distance to the BSs) or because of their coincidental locations in the directions of the beamformed signals. The performance of the proposed fair OUS scheme is demonstrated through simulations in Section VI.

It is necessary to note that the proposed normalization scheme is applicable to any case in which the average channel gain is different for different users, not limited to that caused by path loss. For frequency-selective fading scenarios, multicarrier modulation schemes such as OFDM can be considered and the proposed algorithm, including the channel normalization technique, can be applied to each subcarrier individually.

IV. Group-Rate Balancing for Multigroup Multicasting

In this section, we examine in more detail the GRB problem described in (11) for the multigroup multicasting scenario. The problem is more challenging than that in the single-group scenario because of the existence of cochannel interference among different groups.

Notice that, similar to the single-group scenario, the complexity of solving the GRB problem in (11) is largely due to the combinatorial nature of the search for the optimal user subsets $\{\mathcal{A}_g\}_{g=1}^G$. However, different from the problem in (12), the optimization problem does not reduce to a convex optimization problem even when the user subsets $\{\mathcal{A}_g\}_{g=1}^G$ are given. Specifically, when $\{\mathcal{A}_g\}_{g=1}^G$ are given, we have

$$\begin{aligned}
&\max \quad \min_{g\in\{1,\ldots,G\}} \frac{1}{\tau'_g} R_g && \text{(16a)}\\
&\text{subject to } \log_2\left(1+\frac{\mathrm{tr}(\mathbf{Q}_g\mathbf{h}_k\mathbf{h}_k^H)}{\sum_{\ell\neq g}\mathrm{tr}(\mathbf{Q}_\ell\mathbf{h}_k\mathbf{h}_k^H)+1}\right)\geq R_g,\ \forall k\in\mathcal{A}_g,\forall g, && \text{(16b)}\\
&\sum_{g=1}^G \mathrm{tr}(\{\mathbf{Q}_g\}_{b,b})\leq P_b,\ \forall b,\quad \mathbf{Q}_g\succeq\mathbf{0},\ \forall g, && \text{(16c)}\\
&\{\mathbf{Q}_g\}_{b,b'}=\mathbf{0}_{M\times M},\ \text{for } b\notin\mathcal{B}_g \text{ or } b'\notin\mathcal{B}_g, && \text{(16d)}\\
&\text{variables: } \{\mathbf{Q}_g\}_{g=1}^G, \{R_g\}_{g=1}^G,
\end{aligned}$$

where $\tau'_g \triangleq \tau_g|\mathcal{K}_g|/|\mathcal{A}_g|$. We can observe that the rate variables $\{R_g\}_{g=1}^G$ are unconstrained below and that the objective value does not become smaller by decreasing R_g/τ'_g to the value $\min_{g\in\{1,\ldots,G\}} R_g/\tau'_g$ for all $g=1,\ldots,G$. Consequently, we can impose the constraint

$$R_1/\tau'_1=\cdots=R_G/\tau'_G$$

on problem (16) without loss of optimality.

By introducing the variable $\alpha=R_1/\tau'_1=\cdots=R_G/\tau'_G$, the problem in (16) can be equivalently expressed as

$$\begin{aligned}
&\max \quad \alpha && \text{(17a)}\\
&\text{subject to } \mathrm{tr}(\mathbf{Q}_g\mathbf{h}_k\mathbf{h}_k^H)\geq\left(2^{\tau'_g\alpha}-1\right)\left(\sum_{\ell\neq g}\mathrm{tr}(\mathbf{Q}_\ell\mathbf{h}_k\mathbf{h}_k^H)+1\right),\ \forall k\in\mathcal{A}_g,\forall g, && \text{(17b)}\\
&\sum_{g=1}^G \mathrm{tr}(\{\mathbf{Q}_g\}_{b,b})\leq P_b,\ \forall b,\quad \mathbf{Q}_g\succeq\mathbf{0},\forall g, && \text{(17c)}\\
&\{\mathbf{Q}_g\}_{b,b'}=\mathbf{0}_{M\times M},\ \text{for } b\notin\mathcal{B}_g \text{ or } b'\notin\mathcal{B}_g, && \text{(17d)}\\
&\text{variables: } \{\mathbf{Q}_g\}_{g=1}^G,\ \alpha.
\end{aligned}$$

This problem is still non-convex due to the constraint in (17b). However, for a fixed α, the constraint becomes convex and the problem becomes a convex feasibility problem. The optimal value of α can then be found via a bisection search as described in the following.

To perform the bisection search, it is necessary to first determine an upper bound and a lower bound on the value of α. To do so, notice that, in (17), $\tau_g'\alpha = \tau_g\alpha|\mathcal{K}_g|/|\mathcal{A}_g|$ can be viewed as the minimum rate achievable by all users in the subset $\mathcal{A}_g$ and, thus, is upper-bounded by the rate achievable when all BS powers are used to beamform data to the best user in group g (and no signal is sent to other groups), i.e.,

$$\alpha\frac{\tau_g|\mathcal{K}_g|}{|\mathcal{A}_g|} \le \max_{k\in\mathcal{A}_g}\log_2\left(1+\sum_{b=1}^{B}P_b\|\mathbf{h}_{b,k}\|^2\right), \tag{18}$$

where $\mathbf{h}_{b,k}$ is the channel from BS b to user k. This inequality must hold for all g and, thus, α is upper-bounded by

$$\alpha \le \min_{g\in\{1,\dots,G\}}\max_{k\in\mathcal{A}_g}\frac{|\mathcal{A}_g|}{\tau_g|\mathcal{K}_g|}\log_2\left(1+\sum_{b=1}^{B}P_b\|\mathbf{h}_{b,k}\|^2\right). \tag{19}$$

Moreover, since the rate is non-negative, α is trivially lower-bounded by 0. The bisection algorithm repeatedly bisects the interval between the upper and lower bounds given above until the solution is obtained. Details are given in Algorithm 3.

Algorithm 3 Bisection Search for α

```
1) Initialize $U$ as in (19) and $L = 0$.
2) Set $\alpha = \frac{U+L}{2}$ and solve the feasibility problem corresponding to (17). If it is feasible, then
   set $L = \alpha$. Otherwise, set $U = \alpha$.
3) Repeat Step 2 until $U - L \le \epsilon$, where $\epsilon$ is the value specifying the convergence criteria.
   The resulting $\{\mathbf{Q}_g\}_{g=1}^{G}$ and $\alpha$ are the desired solution to (17).
```

In the above, the optimal transmit covariance matrices $\{\mathbf{Q}_g\}_{g=1}^{G}$ (and the corresponding rates $\{R_g\}_{g=1}^{G}$) were found for given user subsets $\{\mathcal{A}_g\}_{g=1}^{G}$. To find $\{\mathcal{A}_g\}_{g=1}^{G}$, one can employ a heuristic OUS policy similar to the subset search algorithm described in Algorithm 1. However, as mentioned previously, this requires solving (17) in the order of $O(K^2)$ times, with a bisection search embedded in each computation. In the following, we propose a more efficient method based on the introduction of binary user selection variables, as done in the single-group scenario.

Here, an SCA approach is further adopted to cope with the non-convexity caused by the interference terms.

A. Reformulation and Relaxation of the GRB Problem using User Selection Variables

Specifically, following the technique given in Section III, let us introduce the set of binary user selection variables $\{s_k, \forall \mathcal{K}_g, \forall g\}$, where $s_k = 1$ if $k \in \mathcal{A}_g$ for some g, and $s_k = 0$, otherwise. Then, for a sufficiently small δ, the GRB problem in (11) can be equivalently formulated as

$$\max \quad \min_{g\in\{1,\ldots,G\}} \frac{1}{\tau_g|\mathcal{K}_g|} \cdot R_g \sum_{k\in\mathcal{K}_g} s_k \tag{20a}$$

$$\text{subject to } \log_2\left(1+\frac{\operatorname{tr}(\mathbf{Q}_g\mathbf{h}_k\mathbf{h}_k^H)}{\sum_{\ell\neq k}\operatorname{tr}(\mathbf{Q}_\ell\mathbf{h}_k\mathbf{h}_k^H)+1}\right)+\delta^{-1}(1-s_k)\geq R_g,\ \forall k\in\mathcal{K}_g,\ \forall g, \tag{20b}$$

$$\sum_{g=1}^{G}\operatorname{tr}(\{\mathbf{Q}_g\}_{b,b})\leq P_b,\ \forall b, \quad \mathbf{Q}_g\succeq\mathbf{0},\forall g, \tag{20c}$$

$$\{\mathbf{Q}_g\}_{b,b'}=\mathbf{0}_{M\times M},\ \text{for } b\notin\mathcal{B}_g \text{ or } b'\notin\mathcal{B}_g, \tag{20d}$$

$$s_k\in\{0,1\},\ \forall k\in\mathcal{K}, \tag{20e}$$

$$\text{variables: } \{\mathbf{Q}_g\}_{g=1}^G, \{R_g\}_{g=1}^G, \{s_k\}_{k\in\mathcal{K}}.$$

The equivalence of the problems in (11) and (20) is stated in the following lemma.

LEMMA 2: *For*

$$\delta\leq\left[\max_{g\in\mathcal{G}}\max_{k\in\mathcal{K}_g}\log_2(1+\sum_{b=1}^{B}P_b\|\mathbf{h}_{b,k}\|^2)\right]^{-1},$$

$(\{\mathbf{Q}_g^*\}_{g=1}^G, \{R_g^*\}_{g=1}^G, \{s_k^*\}_{k\in\mathcal{K}})$ *is an optimal solution of the problem in* (20) *if and only if* $(\{\mathbf{Q}_g^*\}_{g=1}^G$, $\{R_g^*\}_{g=1}^G, \{\mathcal{A}_g^*\}_{g=1}^G)$, *where* $\mathcal{A}_g^*\triangleq\{k\in\mathcal{K}_g : s_k^*=1\}$, *for* $g=1,\ldots,G$, *is an optimal solution of the GRB problem in* (11).

The proof is similar to that of Lemma 1 and, thus, is omitted. Moreover, let us also consider a relaxation where the integer constraints on s_k's are replaced with the linear constraints $0\leq s_k\leq 1$, for all k. By introducing the auxiliary variable α, the relaxed problem can be written in the epigraph form

$$\begin{aligned} &\max\ \alpha && \text{(21a)}\\ \text{subject to}\ & R_g \sum_{k\in\mathcal{K}_g} s_k \geq \tau_g|\mathcal{K}_g|\alpha^2,\ \forall g, && \text{(21b)}\\ & r_k(\{\mathbf{Q}_\ell\}_{\ell=1}^G) + \delta^{-1}(1-s_k) \geq R_g, \forall k\in\mathcal{K}_g, \forall g, && \text{(21c)}\\ & 0 \leq s_k \leq 1,\ \forall k \in \mathcal{K},\ \text{(20c)},\ \text{and (20d)}, && \text{(21d)}\\ \text{variables:}\ & \{\mathbf{Q}_g\}_{g=1}^G, \{R_g\}_{g=1}^G, \{s_k\}_{k\in\mathcal{K}}, \alpha, \end{aligned}$$

where

$$r_k(\{\mathbf{Q}_\ell\}_{\ell=1}^G) \triangleq \log_2\left(1+\frac{\mathrm{tr}(\mathbf{Q}_g\mathbf{h}_k\mathbf{h}_k^H)}{\sum_{\ell\neq g}\mathrm{tr}(\mathbf{Q}_\ell\mathbf{h}_k\mathbf{h}_k^H)+1}\right), \tag{22}$$

for $k \in \mathcal{K}_g$. Notice that this problem cannot be solved using the bisection search algorithm since, even when α is given, the problem does not reduce to a convex feasibility problem. This is because the value of R_g cannot be determined explicitly and the constraint in (21c) is non-convex. To address this issue, we employ an SCA approach as described in the following.

First, notice that, since R_g is positive, the non-convex constraint in (21b) can be written as

$$\sum_{k\in\mathcal{K}_g} s_k - \alpha\sqrt{\tau_g|\mathcal{K}_g|}R_g^{-1}\alpha\sqrt{\tau_g|\mathcal{K}_g|} \geq 0. \tag{23}$$

Then, by applying the Schur complement [35], this constraint can be equivalently written as the linear matrix inequality constraint

$$\begin{pmatrix} R_g & \alpha\sqrt{\tau_g|\mathcal{K}_g|} \\ \alpha\sqrt{\tau_g|\mathcal{K}_g|} & \sum_{k\in\mathcal{K}_g} s_k \end{pmatrix} \succeq \mathbf{0}. \tag{24}$$

Secondly, to address the non-convexity of the constraint in (21c), we adopt an SCA technique similar to that employed in [36]. Specifically, given any $\{\tilde{\mathbf{Q}}_\ell\}_{\ell=1}^G$ satisfying (20c) and (20d), which together with $\alpha = 0$, $R_g = \delta^{-1}$, $s_k = 0$, $\forall g, k$, is a feasible point to (21), we can rewrite

$r_k(\{\mathbf{Q}_\ell\}_{\ell=1}^G)$ and obtain its lower bound as

$$\begin{aligned} & r_k(\{\mathbf{Q}_\ell\}_{\ell=1}^G) \\ & = \log_2\left(1+\sum_{\ell=1}^{G} \mathrm{tr}(\mathbf{Q}_\ell \mathbf{h}_k \mathbf{h}_k^H)\right) - \log_2\left(1+\sum_{\ell\neq g} \mathrm{tr}(\mathbf{Q}_\ell \mathbf{h}_k \mathbf{h}_k^H)\right) \\ & \geq \log_2\left(1+\sum_{\ell=1}^{G} \mathrm{tr}(\mathbf{Q}_\ell \mathbf{h}_k \mathbf{h}_k^H)\right) - \log_2\left(1+\sum_{\ell\neq g} \mathrm{tr}(\tilde{\mathbf{Q}}_\ell \mathbf{h}_k \mathbf{h}_k^H)\right) - \frac{\sum_{\ell\neq g} \mathrm{tr}((\mathbf{Q}_\ell - \tilde{\mathbf{Q}}_\ell)\mathbf{h}_k \mathbf{h}_k^H)}{[1+\sum_{\ell\neq g} \mathrm{tr}(\tilde{\mathbf{Q}}_\ell \mathbf{h}_k \mathbf{h}_k^H)]\ln 2} \\ & \triangleq \bar{r}_k(\{\mathbf{Q}_\ell\}_{\ell=1}^G \mid \{\tilde{\mathbf{Q}}_\ell\}_{\ell=1}^G) \end{aligned}$$

where the inequality comes from the first-order condition of the concave function $\log_2(\cdot)$, i.e., $\log_2 y \leq \log_2 x + \frac{1}{x \ln 2}(y-x)$ for any $x, y > 0$. Note that $\bar{r}_k(\{\mathbf{Q}_\ell\}_{\ell=1}^G \mid \{\tilde{\mathbf{Q}}_\ell\}_{\ell=1}^G)$ is concave in $\{\mathbf{Q}_\ell\}_{\ell=1}^G$. By replacing $r_k(\{\mathbf{Q}_\ell\}_{\ell=1}^G)$ in constraint (21c) with $\bar{r}_k(\{\mathbf{Q}_\ell\}_{\ell=1}^G \mid \{\tilde{\mathbf{Q}}_\ell\}_{\ell=1}^G)$ and by replacing (21b) with (24), the problem in (21) can be approximated as

$$\max \ \alpha \tag{25a}$$

$$\text{subject to} \ \begin{pmatrix} R_g & \alpha\sqrt{\tau_g|\mathcal{K}_g|} \\ \alpha\sqrt{\tau_g|\mathcal{K}_g|} & \sum_{k\in\mathcal{K}_g} s_k \end{pmatrix} \succeq \mathbf{0}, \ \forall g, \tag{25b}$$

$$\bar{r}_k(\{\mathbf{Q}_\ell\}_{\ell=1}^G \,|\, \{\tilde{\mathbf{Q}}_\ell\}_{\ell=1}^G) + \delta^{-1}(1-s_k) \geq R_g, \ \forall k \in \mathcal{K}_g, \ \forall g, \tag{25c}$$

$$0 \leq s_k \leq 1, \ \forall k \in \mathcal{K}, \ \text{(20c), and (20d)}, \tag{25d}$$

$$\text{variables: } \{\mathbf{Q}_g\}_{g=1}^G, \{R_g\}_{g=1}^G, \{s_k\}_{k\in\mathcal{K}}, \alpha.$$

The resulting optimization problem in (25) is convex and can be solved using general purpose solvers such as CVX [34]. Notice that this approximation is conservative in the sense that solving (25) yields a feasible approximate solution to (21). However, the approximation performance can be improved by iteratively applying the same approximation to (21), with $\{\tilde{\mathbf{Q}}_\ell\}_{\ell=1}^G$ taken as the solution of $\{\mathbf{Q}_\ell\}_{\ell=1}^G$ obtained in the previous iteration. The process can be repeated until the value of α converges. The algorithm is summarized in Algorithm 4.

Algorithm 4 Relaxed GRB via Successive Convex Approximation (SCA)

```
1) Initialize $\{\tilde{\mathbf{Q}}_\ell\}_{\ell=1}^G$ by any feasible point of (20c), (20d), and set $\tilde{\alpha} = 0$.
2) Solve problem (25), and denote the solution of $\mathbf{Q}_\ell$ as $\mathbf{Q}_\ell^*$ and the objective value as $\alpha^*$.
3) If $|\tilde{\alpha} - \alpha^*| \leq \epsilon$, then stop; else, set $\tilde{\mathbf{Q}}_\ell \leftarrow \mathbf{Q}_\ell^*$ for all $\ell$, $\tilde{\alpha} \leftarrow \alpha^*$, and go to Step 2.
```

The convergence of Algorithm 4 is described in the following proposition.

PROPOSITION 1: *Let* $\{\{\mathbf{Q}_g^*\}_{g=1}^G, \{R_g^*\}_{g=1}^G, \{s_k^*\}_{k\in\mathcal{K}}, \alpha^*\}$ *denote the solution of* (25). *Then, any limit point of the iterates* $\{\{\mathbf{Q}^*\}_{g=1}^G, \{R_g^*\}_{g=1}^G, \{s_k^*\}_{k\in\mathcal{K}}, \alpha^*\}$ *generated by Algorithm 4 is a stationary point of problem* (21).

The proof is given in Appendix B. After obtaining the values of s_k from the relaxed problem (and the SCA approach in Algorithm 4), the sequential deflation technique proposed in the previous section can then be utilized to obtain a solution for the user subsets $\{\mathcal{A}_g\}_{g=1}^G$. Similarly, by employing the sequential deflation technique, the optimization problem in (25) is solved $O(K)$ times, which is a significant improvement over the subset search with bisection. The performance of the two algorithms will be compared in Section VI.

B. Fairness of the GRB Problem in the Non-I.I.D. Case based on Channel Normalization

Notice that, similar to the single-group scenario, the GRB problem may also suffer fairness issues since the average group-rate is again used as the maximization criterion in each group. In this section, we extend the heuristic channel normalization technique, previously proposed in Section III-B, to the multi-group scenario.

Specifically, suppose that the channel vector $\mathbf{h}_{b,k}[n]$ between BS b and user k in block n has entries that are i.i.d. $\mathcal{CN}(0, d_{b,k}^{-\alpha})$, where $d_{b,k}$ is the distance between BS b and user k and α is the path loss exponent. Let $\bar{d}_{b,g} = \frac{1}{|\mathcal{K}_g|}\sum_{k\in\mathcal{K}_g} d_{b,k}$ be the average distance between BS b and users in group g. Then, the normalized channel vector between BS b and user k in group g can be defined as

$$\tilde{\mathbf{h}}_{b,k}[n] = \mathbf{h}_{b,k}[n]\sqrt{\frac{\bar{d}_{b,g}^{-\alpha}}{d_{b,k}^{-\alpha}}}. \tag{26}$$

Here, instead of scaling the normalized channel vectors by the average distance associated with all users, as in (26), the channel vectors are scaled by the average distance of users within its own group. Similar to Section III-B, the normalized channel vectors are then utilized to compute the optimal user subsets, denoted by $\tilde{\mathcal{A}}_g^*$, for all g, using the algorithm proposed in the previous subsection. With user subsets $\{\tilde{\mathcal{A}}^*\}_{g=1}^G$, the optimal transmit covariance matrix can then be computed by solving (16) using the original channel vectors $\mathbf{h}_{b,k}$, $\forall b, k$.

By performing OUS based on the normalized channel vectors, users in the same group will have equal probability of being selected in each block. These users will likely be those whose instantaneous channels are temporarily more favorable, either in the sense that their channel gains are large relative to their respective averages or in the sense that their channel directions cause less interference to other groups. The performance of the proposed fair OUS scheme is demonstrated through simulations in Section VI.

V. Group-Rate Balancing based on Partial Channel Feedback

In previous sections, the design of the transmit covariance matrices and the user subsets was performed based on instantaneous knowledge of all users' CSI. In practice, this may require a large amount of feedback, especially when the number of users in the system is large. In this section, we consider the case where only a subset of users feeds back their CSI in each block. Since the CSI of some users may be missing, the optimization can only be performed based on the expected group-rate, i.e., the expectation of the average group-rate with respect to the statistics of the unknown channels.

Let $\mathcal{K}_{g,\text{fb}}$ be the set of users from group $\mathcal{K}_g$ that have chosen to feedback their channels in the given block and let $\mathcal{K}_{g,\text{fb}}^c = \mathcal{K}_g - \mathcal{K}_{g,\text{fb}}$ be the complement set, i.e., the set of users from group $\mathcal{K}_g$ whose CSI is unknown. In this case, the system parameters can be found by solving the group-rate balancing with partial feedback (GRB-PF) problem as given below:

GRB with Partial Feedback (GRB-PF) Problem:

$$\max \ \min_g \frac{R_g}{\tau_g|\mathcal{K}_g|}\mathbf{E}\left[\sum_{k\in\mathcal{K}_g}\mathbf{1}_{\left\{r_k(\{\mathbf{Q}_\ell\}_{\ell=1}^G)\geq R_g\right\}}\right] \tag{27a}$$

$$\text{subject to } \sum_{g=1}^{G}\text{tr}(\{\mathbf{Q}_g\}_{b,b}) \leq P_b, \ \forall b, \quad \mathbf{Q}_g \succeq \mathbf{0}, \forall g, \tag{27b}$$

$$\{\mathbf{Q}_g\}_{b,b'} = \mathbf{0}_{M\times M}, \text{ for } b \notin \mathcal{B}_g \text{ or } b' \notin \mathcal{B}_g, \tag{27c}$$

$$\text{variables: } \{\mathbf{Q}_g\}_{g=1}^G, \{R_g\}_{g=1}^G,$$

Here, we assume that a user's decision to feedback its channel (or not) is independent of its channel realizations. Otherwise, the expectation must be evaluated by conditioning on the specific feedback strategy[1]. In particular, the expectation term inside the objective function can be written as

$$\sum_{k\in\mathcal{K}_{g,\text{fb}}} \mathbf{1}_{\left\{\log_2\left(1+\frac{\text{tr}(\mathbf{Q}_g\mathbf{h}_k\mathbf{h}_k^H)}{\sum_{\ell\neq g}\text{tr}(\mathbf{Q}_\ell\mathbf{h}_k\mathbf{h}_k^H)+1}\right)\geq R_g\right\}} + \sum_{k\in\mathcal{K}^c_{g,\text{fb}}} \Pr\left(\log_2\left(1+\frac{\text{tr}(\mathbf{Q}_g\mathbf{h}_k\mathbf{h}_k^H)}{\sum_{\ell\neq g}\text{tr}(\mathbf{Q}_\ell\mathbf{h}_k\mathbf{h}_k^H)+1}\right)\geq R_g\right). \tag{28}$$

Notice that the probability in the second term cannot be evaluated in closed-form and, thus, is difficult to handle from an optimization perspective. To address this issue, we propose to consider an approximation technique where the probability is replaced with its sample average approximation (SAA). Similar techniques have also been widely adopted in the area of stochastic optimization, e.g., [26], [27].

More specifically, let $f_{\mathbf{h}_k}$ be the density function of $\mathbf{h}_k$ and let $\{\mathbf{h}_k^{(j)}\}_{j=1}^J$ be J vectors randomly generated according to $f_{\mathbf{h}_k}$. Then, the probability can be approximated as

$$\Pr\left(\log_2\left(1+\frac{\text{tr}(\mathbf{Q}_g\mathbf{h}_k\mathbf{h}_k^H)}{\sum_{\ell\neq g}\text{tr}(\mathbf{Q}_\ell\mathbf{h}_k\mathbf{h}_k^H)+1}\right)\geq R_g\right) \approx \frac{1}{J}\sum_{j=1}^{J} \mathbf{1}_{\left\{\log_2\left(1+\frac{\text{tr}(\mathbf{Q}_g\mathbf{h}_k^{(j)}(\mathbf{h}_k^{(j)})^H)}{\sum_{\ell\neq g}\text{tr}(\mathbf{Q}_\ell\mathbf{h}_k^{(j)}(\mathbf{h}_k^{(j)})^H)+1}\right)\geq R_g\right\}}. \tag{29}$$

With knowledge of the distributions of the channel vectors $\mathbf{h}_k$, for each $k\in\mathcal{K}^c_{g,\text{fb}}$, the objective function in (27) can be approximated as

$$\min_{g}\ \frac{R_g}{\tau_g|\mathcal{K}_g|}\left[\sum_{k\in\mathcal{K}_{g,\text{fb}}}\mathbf{1}_{\left\{r_k(\{\mathbf{Q}_\ell\}_{\ell=1}^G)\geq R_g\right\}}+\sum_{k\in\mathcal{K}^c_{g,\text{fb}}}\frac{1}{J}\sum_{j=1}^{J}\mathbf{1}_{\left\{r_{k,j}(\{\mathbf{Q}_\ell\}_{\ell=1}^G)\geq R_g\right\}}\right], \tag{30}$$

where

$$r_{k,j}(\{\mathbf{Q}_\ell\}_{\ell=1}^G)\triangleq\log_2\left(1+\frac{\text{tr}(\mathbf{Q}_g\mathbf{h}_k^{(j)}(\mathbf{h}_k^{(j)})^H)}{\sum_{\ell\neq g}\text{tr}(\mathbf{Q}_\ell\mathbf{h}_k^{(j)}(\mathbf{h}_k^{(j)})^H)+1}\right), \tag{31}$$

for $k\in\mathcal{K}^c_{g,\text{fb}}$. The channel vectors $\{\mathbf{h}_k^{(j)}\}_{j=1}^J$ can be viewed as the channel vectors of J virtual users associated with user k. The entire set of virtual users in group g is defined as $\mathcal{K}_{g,\text{vir}}\triangleq\mathcal{K}^c_{g,\text{fb}}\times\{1,\ldots,J\}$. Moreover, let $\mathcal{A}_{g,\text{fb}}\subseteq\mathcal{K}_{g,\text{fb}}$ and $\mathcal{A}_{g,\text{vir}}\subseteq\mathcal{K}_{g,\text{vir}}$ be the subsets of selected feedback and virtual users, respectively, in group g that are intended to successfully decode, and

[1] Note that, in cases with partial feedback (i.e., when only part of the users feedback their channels), the specific feedback strategy may have a significant impact on the group-rate performance. In particular, allowing channel feedback from users with bad channels may improve service towards these users in each block but may limit the achievable multiuser and temporal diversity gains over time; and allowing feedback from users with good channels may have the opposite effect. The optimal feedback strategy should exploit the tradeoff between these two effects, especially in the non-i.i.d. case, but requires further studies that go beyond the scope of this work. Readers are referred to [37] and [38] for further discussions on this topic.

let the index pair (k, j) denote the j-th virtual user associated with (non-feedback) user k. The GRB-PF problem in (27) can then be approximated as follows:

$$\begin{aligned} \max \quad & \min_{g\in\{1,\ldots,G\}} \frac{R_g}{\tau_g|\mathcal{K}_g|}\left(|\mathcal{A}_{g,\mathrm{fb}}| + \frac{1}{J}|\mathcal{A}_{g,\mathrm{vir}}|\right) && \text{(32a)}\\ \text{subject to } & r_k(\{\mathbf{Q}_\ell\}_{\ell=1}^G) \geq R_g,\ \forall k \in \mathcal{A}_{g,\mathrm{fb}}, \forall g, && \text{(32b)}\\ & r_{k,j}(\{\mathbf{Q}_\ell\}_{\ell=1}^G) \geq R_g,\ \forall (k,j) \in \mathcal{A}_{g,\mathrm{vir}}, \forall g, && \text{(32c)}\\ & \text{(27b) and (27c)}, && \text{(32d)}\\ \text{variables: } & \{\mathbf{Q}_g\}_{g=1}^G, \{R_g\}_{g=1}^G, \{\mathcal{A}_{g,\mathrm{fb}}\}_{g=1}^G, \{\mathcal{A}_{g,\mathrm{vir}}\}_{g=1}^G. \end{aligned}$$

Notice that the above problem is difficult to solve due to the combinatorial nature of the search over the user subsets $\{\mathcal{A}_{g,\mathrm{fb}}\}_{g=1}^G$ and $\{\mathcal{A}_{g,\mathrm{vir}}\}_{g=1}^G$. However, when $\{\mathcal{A}_{g,\mathrm{fb}}\}_{g=1}^G$ and $\{\mathcal{A}_{g,\mathrm{vir}}\}_{g=1}^G$ are given, the problem becomes the same as (16) with $\tau'_g \triangleq \tau_g|\mathcal{K}_g|/(|\mathcal{A}_{g,\mathrm{fb}}| + \frac{1}{J}|\mathcal{A}_{g,\mathrm{vir}}|)$ and, thus, can be solved using the bisection search algorithm described in Algorithm 3.

To determine the user selection (i.e., the choice of $\{\mathcal{A}_{g,\mathrm{fb}}\}_{g=1}^G$ and $\{\mathcal{A}_{g,\mathrm{vir}}\}_{g=1}^G$), we introduce, for each feedback user $k \in \mathcal{K}_{g,\mathrm{fb}}$, a user selection variable s_k and, for each virtual user $(k, j) \in \mathcal{K}_{g,\mathrm{vir}}$, a user selection variable $t_{k,j}$. Then, similar to the previous sections, the approximated GRB-PF problem can be reformulated as

$$\begin{aligned} \max \quad & \min_{g} \frac{R_g}{\tau''_g}\left(\sum_{k\in\mathcal{K}_{g,\mathrm{fb}}} s_k + \frac{1}{J}\sum_{(k,j)\in\mathcal{K}_{g,\mathrm{vir}}} t_{k,j}\right) && \text{(33a)}\\ \text{subject to } & r_k(\{\mathbf{Q}_\ell\}_{\ell=1}^G) + \delta^{-1}(1 - s_k) \geq R_g,\ \forall k \in \mathcal{K}_{g,\mathrm{fb}},\ \forall g, && \text{(33b)}\\ & r_{k,j}(\{\mathbf{Q}_\ell\}_{\ell=1}^G) + \delta^{-1}(1 - t_{k,j}) \geq R_g,\ \forall (k,j) \in \mathcal{K}_{g,\mathrm{vir}}, \forall g, && \text{(33c)}\\ & s_k \in \{0,1\},\ \forall k \in \mathcal{K}_{g,\mathrm{fb}}, \forall g,\ \text{(27b)}, \text{(27c)}, && \text{(33d)}\\ & t_{k,j} \in \{0,1\},\ \forall (k,j) \in \mathcal{K}_{g,\mathrm{vir}},\ \forall g, && \text{(33e)}\\ \text{variables: } & \{\mathbf{Q}_g\}_{g=1}^G, \{R_g\}_{g=1}^G, \{s_k, \forall k \in \mathcal{K}_{g,\mathrm{fb}}\}_{g=1}^G, \{t_{k,j}, \forall (k,j) \in \mathcal{K}_{g,\mathrm{vir}}\}_{g=1}^G. \end{aligned}$$

where $\tau''_g \triangleq \tau_g|\mathcal{K}_g|$.

By relaxing the integer constraints and by applying properties of the Schur complement, we can obtain a similar relaxed problem in its epigraph form, i.e.,

$$\max \ \alpha \tag{34a}$$

$$\text{subject to } \begin{pmatrix} R_g & \alpha\sqrt{\tau''_g} \\ \alpha\sqrt{\tau''_g} & \sum\limits_{k\in\mathcal{K}_g} s_k + \frac{1}{J}\sum_{(k,j)\in\mathcal{K}_{g,\text{vir}}} t_{k,j} \end{pmatrix} \succeq \mathbf{0},\ \forall g, \tag{34b}$$

$$r_k(\{\mathbf{Q}_\ell\}_{\ell=1}^G) + \delta^{-1}(1-s_k) \geq R_g,\ \forall k \in \mathcal{K}_{g,\text{fb}},\ \forall g, \tag{34c}$$

$$r_{k,j}(\{\mathbf{Q}_\ell\}_{\ell=1}^G) + \delta^{-1}(1-t_{k,j}) \geq R_g,\ \forall (k,j) \in \mathcal{K}_{g,\text{vir}},\ \forall g, \tag{34d}$$

$$0 \leq s_k \leq 1,\ \forall k \in \mathcal{K}_{g,\text{fb}},\ \text{(27b)}, \text{(27c)}, \tag{34e}$$

$$0 \leq t_{k,j} \leq 1,\ \forall (k,j) \in \mathcal{K}_{g,\text{vir}},\ \forall g, \tag{34f}$$

$$\text{variables: } \{\mathbf{Q}_g\}_{g=1}^G, \{R_g\}_{g=1}^G, \{s_k, \forall k \in \mathcal{K}_{g,\text{fb}}\}_{g=1}^G, \{t_{k,j}, \forall (k,j) \in \mathcal{K}_{g,\text{vir}}\}_{g=1}^G, \alpha.$$

The above problem is still non-convex due to the constraints in (34c) and (34d). Therefore, we adopt an SCA approach, similar to that in the previous section, where the left-hand-side of the inequalities in (34c) and (34d) are approximated by their concave lower bounds. More specifically, the problem in (34) is approximated as

$$\max \ \alpha \tag{35a}$$

$$\text{subject to } \begin{pmatrix} R_g & \alpha\sqrt{\tau''_g} \\ \alpha\sqrt{\tau''_g} & \sum\limits_{k\in\mathcal{K}_g} s_k + \frac{1}{J}\sum_{(k,j)\in\mathcal{K}_{g,\text{vir}}} t_{k,j} \end{pmatrix} \succeq \mathbf{0},\ \forall g, \tag{35b}$$

$$\bar{r}_k(\{\mathbf{Q}_\ell\}_{\ell=1}^G \mid \{\tilde{\mathbf{Q}}_\ell\}_{\ell=1}^G) + \delta^{-1}(1-s_k) \geq R_g,\ \forall k \in \mathcal{K}_{g,\text{fb}},\ \forall g, \tag{35c}$$

$$\bar{r}_{k,j}(\{\mathbf{Q}_\ell\}_{\ell=1}^G \mid \{\tilde{\mathbf{Q}}_\ell\}_{\ell=1}^G) + \delta^{-1}(1-t_{k,j}) \geq R_g,\ \forall (k,j) \in \mathcal{K}_{g,\text{vir}},\ \forall g, \tag{35d}$$

$$0 \leq s_k \leq 1,\ \forall k \in \mathcal{K}_{g,\text{fb}},\ \text{(27b)}, \text{(27c)}, \tag{35e}$$

$$0 \leq t_{k,j} \leq 1,\ \forall (k,j) \in \mathcal{K}_{g,\text{vir}},\ \forall g, \tag{35f}$$

$$\text{variables: } \{\mathbf{Q}_g\}_{g=1}^G, \{R_g\}_{g=1}^G, \{s_k, \forall k \in \mathcal{K}_{g,\text{fb}}\}_{g=1}^G, \{t_{k,j}, \forall (k,j) \in \mathcal{K}_{g,\text{vir}}\}_{g=1}^G, \alpha,$$

where $\{\tilde{\mathbf{Q}}_\ell\}_{\ell=1}^G$ can be any point satisfying constraints (27b) and (27c), and $\bar{r}_{k,j}(\{\mathbf{Q}_\ell\}_{\ell=1}^G \mid$

$\{\tilde{\mathbf{Q}}_\ell\}_{\ell=1}^G)$ is defined as

$$\bar{r}_{k,j}(\{\mathbf{Q}_\ell\}_{\ell=1}^G \mid \{\tilde{\mathbf{Q}}_\ell\}_{\ell=1}^G) \triangleq \log_2\left(1+\sum_{\ell=1}^G \mathrm{tr}(\mathbf{Q}_\ell \mathbf{h}_k^{(j)}(\mathbf{h}_k^{(j)})^H)\right) - \log_2\left(1+\sum_{\ell\neq g}\mathrm{tr}(\tilde{\mathbf{Q}}_\ell \mathbf{h}_k^{(j)} \mathbf{h}_k^{(j)^H})\right)$$
$$- \frac{\sum_{\ell\neq g}\mathrm{tr}((\mathbf{Q}_\ell - \tilde{\mathbf{Q}}_\ell)\mathbf{h}_k^{(j)}\mathbf{h}_k^{(j)^H})}{\left[1+\sum_{\ell\neq g}\mathrm{tr}(\tilde{\mathbf{Q}}_\ell \mathbf{h}_k^{(j)}\mathbf{h}_k^{(j)^H})\right]\ln 2}.$$

The resulting optimization problem (35) is convex and can be solved using general purpose solvers such as CVX [34]. Similarly, refined approximate solutions can be obtained by solving a series of convex optimization problems where, in each iteration, $\{\tilde{\mathbf{Q}}_\ell\}_{\ell=1}^G$ are taken as the solutions of $\{\mathbf{Q}_\ell\}_{\ell=1}^G$ obtained in the previous iteration. The process can be repeated until the objective value α converges. The algorithm is summarized in Algorithm 5.

Algorithm 5 Relaxed GRB-PF via SAA and SCA

```
1) Randomly generate channel vectors $\{\mathbf{h}_k^{(j)}\}_{j=1}^J, \forall k \in \mathcal{K}_{g,\mathrm{fb}}^c, \forall g$.
2) Initialize $\{\tilde{\mathbf{Q}}_\ell\}_{\ell=1}^G$ by any feasible point of (27b), (27c), and set $\tilde{\alpha} = 0$.
3) Solve problem (35), and denote the solution of $\mathbf{Q}_\ell$ as $\mathbf{Q}_\ell^*$ , and the objective value as $\alpha^*$.
4) If $|\tilde{\alpha} - \alpha^*| \leq \epsilon$, then stop; else, set $\tilde{\mathbf{Q}}_\ell \leftarrow \mathbf{Q}_\ell^*$, $\tilde{\alpha} \leftarrow \alpha^*$, and go to Step 2.
```

PROPOSITION 2: *Let* $(\{\mathbf{Q}_g^*\}_{g=1}^G, \{R_g^*\}_{g=1}^G, \alpha^*, \{s_k^*, \forall k \in \mathcal{K}_{g,\mathrm{fb}}\}_{g=1}^G, \{t_{k,j}^*, \forall (k,j) \in \mathcal{K}_{g,\mathrm{vir}}\}_{g=1}^G)$ *denote the solution of* (35)*. Then, any limit point of the iterates generated by Algorithm 5 is a stationary point of problem* (34)*.*

The proof of Proposition 2 is similar to that of Proposition 1 and, thus, is omitted. After obtaining the values of s_k and $t_{k,j}$ from the relaxed problem, a sequential deflation technique, similar to that in Algorithm 2, is again applied to obtain a solution for the feedback and virtual user subsets, i.e., $\{\mathcal{A}_{g,\mathrm{fb}}^*\}_{g=1}^G$ and $\{\mathcal{A}_{g,\mathrm{vir}}^*\}_{g=1}^G$, respectively. The sequential deflation algorithm used for the GRB-PF problem is formally described in Algorithm 6. Notice that, in this problem, SAA introduces a large number of virtual users which may increase the complexity of the sequential deflation approach. To reduce complexity, we choose to eliminate more than one virtual user at once in each iteration. Specifically, if the smallest user selection variable is associated with a virtual user (i.e., if $\min_{(k,j)\in\mathcal{A}_{g,\mathrm{vir}}} t_{k,j} \leq \min\{\min_{k\in\cup_g\mathcal{A}_{g,\mathrm{fb}}} s_k, \min_{(k,j)\in\cup_g\mathcal{A}_{g,\mathrm{vir}}} t_{k,j}\}$), then $D_t > 1$ (e.g., $D_t = 5$) virtual users with the D_t smallest user selection variables are eliminated from the potential virtual user subset $\mathcal{A}_{g,\mathrm{vir}}$. On the other hand, if $\min_{k\in\mathcal{A}_{g,\mathrm{fb}}} s_k \leq \min\{\min_{k\in\cup_g\mathcal{A}_{g,\mathrm{fb}}} s_k, \min_{(k,j)\in\cup_g\mathcal{A}_{g,\mathrm{vir}}} t_{k,j}\}$, then $D_s = 1$ feedback user is eliminated from the potential feedback user subset $\mathcal{A}_{g,\mathrm{fb}}$. The removal of virtual users can be viewed as the removal of channel realizations that are not able to jointly support a high transmission rate.

Algorithm 6 OUS by Sequential Deflation for the GRB-PF Problem

```
(i)   Initialize by setting $\mathcal{A}_{g,\mathrm{fb}} \leftarrow \mathcal{K}_{g,\mathrm{fb}}$, $\mathcal{A}_{g,\mathrm{vir}} \leftarrow \mathcal{K}_{g,\mathrm{vir}}$, $\mathcal{A}^*_{g,\mathrm{fb}} \leftarrow \emptyset$, $\mathcal{A}^*_{g,\mathrm{vir}} \leftarrow \emptyset$, for $g = 1, \ldots, G$, and $\alpha^* \leftarrow 0$.
(ii)  Solve (32) for given user subsets $\{\mathcal{A}_{g,\mathrm{fb}}\}_{g=1}^G$ and $\{\mathcal{A}_{g,\mathrm{vir}}\}_{g=1}^G$, and let $\tilde{\alpha}$ be the resulting objective value.
(iii) If $\tilde{\alpha} > \alpha^*$, then set $\alpha^* \leftarrow \tilde{\alpha}$, $\mathcal{A}^*_{g,\mathrm{fb}} \leftarrow \mathcal{A}_{g,\mathrm{fb}}$, and $\mathcal{A}^*_{g,\mathrm{vir}} \leftarrow \mathcal{A}_{g,\mathrm{vir}}$, $\forall g$.
(iv)  Solve the relaxed GRB-PF problem in (35) (using Algorithm 5) for $\mathcal{K}_{g,\mathrm{fb}} = \mathcal{A}_{g,\mathrm{fb}}$ and $\mathcal{K}_{g,\mathrm{vir}} = \mathcal{A}_{g,\mathrm{vir}}$, $\forall g$, to yield the values of $s_k$, $\forall k \in \mathcal{A}_{g,\mathrm{fb}}$, and $t_{k,j}$, $\forall (k,j) \in \mathcal{A}_{g,\mathrm{vir}}$, $\forall g$.
(v)   If $\min_{k\in\mathcal{A}_{g,\mathrm{fb}}} s_k \leq \min\left\{\min_{k\in\cup_g\mathcal{A}_{g,\mathrm{fb}}} s_k,\ \min_{(k,j)\in\cup_g\mathcal{A}_{g,\mathrm{vir}}} t_{k,j}\right\}$, then set $\mathcal{A}_{g,\mathrm{fb}} \leftarrow \mathcal{A}_{g,\mathrm{fb}} - \mathcal{S}_{\min,\mathrm{fb}}$, where $\mathcal{S}_{\min,\mathrm{fb}}$ is the set of feedback users associated with the $D_s$ smallest values of $s_k$ among users in $\mathcal{A}_{g,\mathrm{fb}}$; else, if $\min_{(k,j)\in\mathcal{A}_{g,\mathrm{vir}}} t_{k,j} \leq \min\left\{\min_{k\in\cup_g\mathcal{A}_{g,\mathrm{fb}}} s_k, \min_{(k,j)\in\cup_g\mathcal{A}_{g,\mathrm{vir}}} t_{k,j}\right\}$, then set $\mathcal{A}_{g,\mathrm{vir}} \leftarrow \mathcal{A}_{g,\mathrm{vir}} - \mathcal{S}_{\min,\mathrm{vir}}$, where $\mathcal{S}_{\min,\mathrm{vir}}$ is the set of virtual users associated with the $D_t$ smallest values of $t_{k,j}$ in $\mathcal{A}_{g,\mathrm{vir}}$.
(vi)  Repeat steps (ii)-(v) until $\mathcal{A}_{g,\mathrm{fb}} = \emptyset$ for some $g$. Then, take $\{\mathcal{A}^*_{g,\mathrm{fb}}\}_{g=1}^G$ and $\{\mathcal{A}^*_{g,\mathrm{vir}}\}_{g=1}^G$ as the desired set of feedback and virtual users.
```

It is worthwhile to note that, in the proposed scheme, the virtual users are generated at random and, thus, the solution obtained under certain realizations of the virtual users' channels may be worse than the case without virtual users (i.e., solving the GRB problem assuming only the existence of the feedback users). This is especially the case when the number of virtual users, i.e., J, is small. Hence, in practice, one can take the better of the two schemes (i.e., the scheme considering only feedback users' channels and the proposed GRB-PF scheme) in each realization to guarantee that no loss occurs due to such randomness. This allows us to reduce the required number of virtual users in practice. Moreover, when the users' channels are i.i.d., the same set of virtual users can be used for all users, which further reduces the complexity.

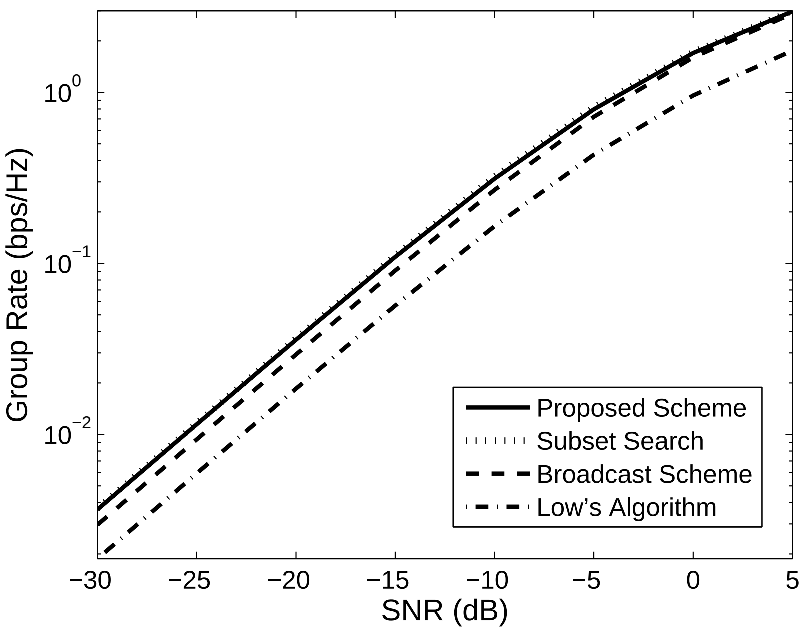

Fig. 3. For the single-group multicasting scenario with $B = 3$ BSs, $M = 2$ antennas per BS, and $K = 30$ users, we show the average group-rates of single-group multicasting using the proposed sequential deflation, the subset search, the broadcast, and Low’s algorithms.

VI. Simulations Results

In this section, the effectiveness of the proposed schemes is demonstrated through computer simulations. In the experiments, we consider a multicell network with $B = 3$ BSs, each equipped with $M = 2$ transmit antennas, and set the power constraints of the BSs as $P_1 = P_2 = P_3 = P$. The SNR is defined as P/σ^2, where $\sigma^2 = 1$ is the noise variance at each receiver (as chosen in Section II), and the entries of the channel vectors are assumed to be i.i.d. $\mathcal{CN}(0, 1)$ unless mentioned otherwise. The results are obtained by averaging over 600 channel realizations.

A. Single-Group Multicasting Scenario

First, we consider the single-group multicasting scenario with $B = 3$ BSs serving collaboratively $K = 30$ users. We would like to emphasize that the scenario under consideration is different from having a single BS with 6 antennas since each BS here is subject to their own individual power constraint P. In Fig. 3, we compare the average group-rate of single-group multicasting using the proposed sequential deflation technique (described in Algorithm 2) with that of single-group multicasting using the subset search algorithm (described in Algorithm 1), the broadcast

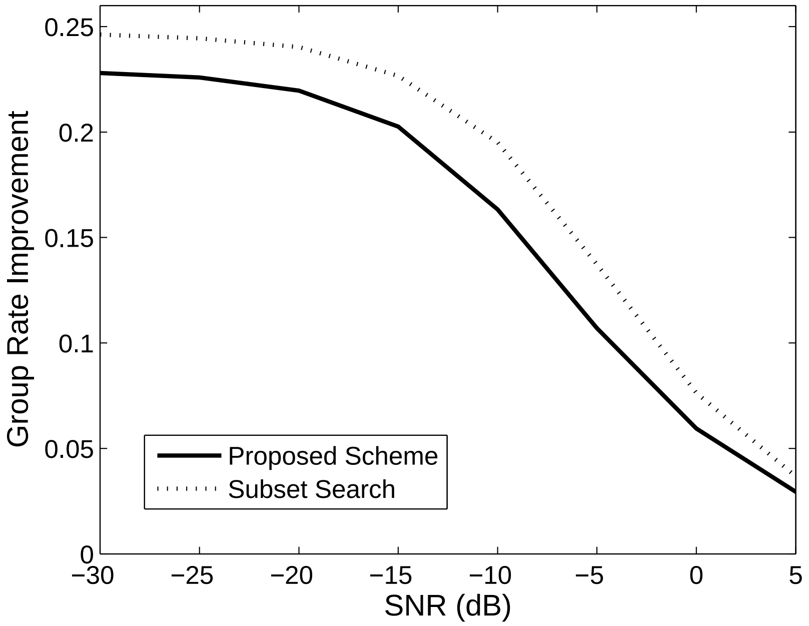

Fig. 4. For the single-group multicasting scenario with $B = 3$ BSs, $M = 2$ antennas per BS, and $K = 30$ users, we show the average group-rate improvements obtained with the proposed sequential deflation algorithm and the subset search algorithm.

scheme (where all users are served in each block), and Low's algorithm [21] (which is based on a heuristic semi-orthogonal user selection algorithm). Here, Low's algorithm is performed with power allocation that takes into consider the individual power constraints at different BSs. We can see that, in the single group scenario, the proposed and the subset-search based OUS policies perform better than the broadcast scheme and, in fact, provide more advantages in the low SNR regime than in the high SNR regime. This is because, in the high SNR regime, a significant rate loss may be experienced when a user is eliminated and, thus, it is preferable to serve all users simultaneously (as done in the broadcast scheme). In cases with OUS, the subset search algorithm performs slightly better than the proposed scheme, but the difference is not significant and comes at the cost of much higher complexity. Low's algorithm performs the worst among all scehems since the precoder is not shaped in accordance with the individual power constraints, but chosen to maintain orthogonality among different signal directions [21]. In Fig. 4, we show the group-rate improvement of the proposed and the subset search algorithms. The group-rate improvement is defined as the difference in group-rate between the compared algorithm and the broadcast scheme, normalized by the group-rate of the latter scheme. We can see that, at low

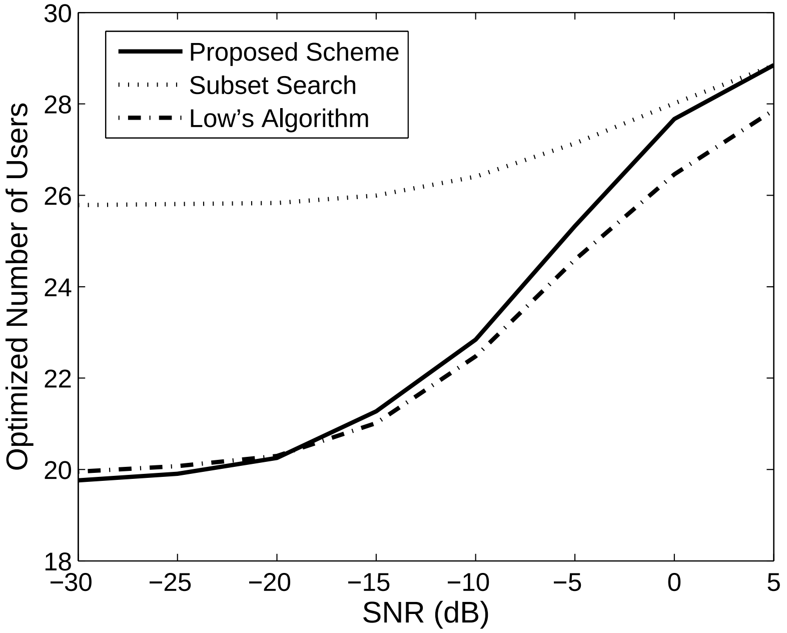

Fig. 5. For the single-group multicasting scenario with $B = 3$ BSs, $M = 2$ antennas per BS, and $K = 30$ users, we show the average number of selected users obtained with the proposed sequential deflation, the subset search, and Low's algorithms.

SNR, the group-rate improvement can be over 20% for both the proposed and the subset search algorithms.

In Fig. 5, we show the average number of users that are selected in each block in the proposed sequential deflation, the subset search, and Low's algorithms. We can see that the number of selected users increases with SNR in all schemes. The subset search algorithm eliminates fewer users because it terminates whenever no further improvement is obtained after removing a user whereas the proposed scheme first removes users sequentially until no user remains to yield $|\mathcal{K}|$ candidate user subsets and then chooses the solution that yields the best group-rate. Even though the subset search algorithm is able to choose the most appropriate user to eliminate in each iteration (since it performs an exhaustive search among all remaining users in each iteration), it may have terminated prematurely because of the existence of many locally optimum solutions. It is worthwhile to note that the number of users selected in each block does not reflect the long-term fairness of the scheme. It only shows how each scheme exploits the tradeoff between multiuser diversity and multicast gains. When the users's channels are i.i.d., each user has equal opportunity of being selected in each block and thus, with the incorporation of the outer code,

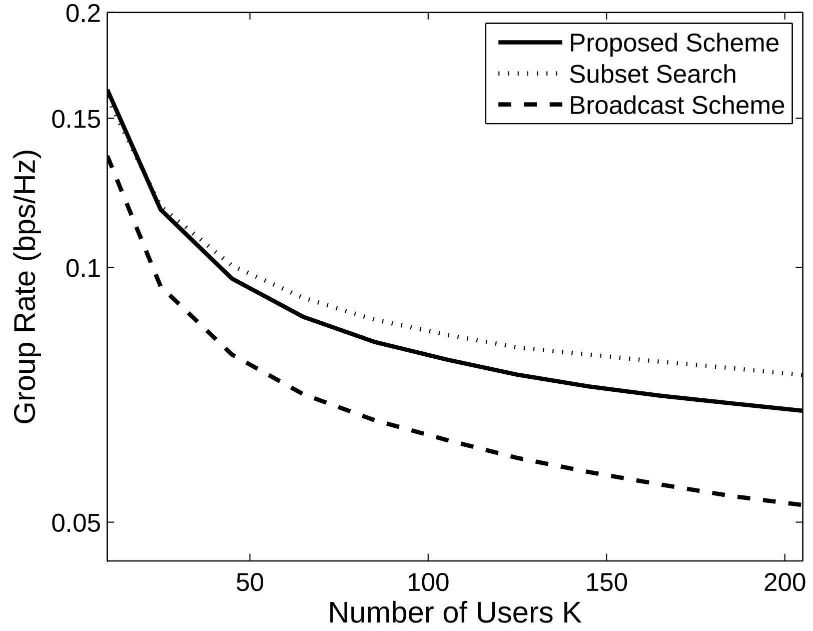

Fig. 6. For the single-group multicasting scenario with $B = 3$ BSs, $M = 2$ antennas per BS, and SNR=-15dB, we show the average group-rate of the proposed, the subset search, and the broadcast schemes as the number of users K increases.

the average rate achieved by all users is asymptotically the same.

In Fig. 6, we show the average group-rate of the proposed, the subset search, and the broadcast schemes with respect to the number of users, i.e., K. The receive SNR is fixed as -15dB. Due to the high computational complexity of the subset search scheme, its performance is averaged only over 300 channel realizations for $K \leq 125$ and 30 channel realizations for $K > 125$. We can see that the OUS schemes (i.e., the proposed and the subset search algorithms) perform better than the broadcast scheme, especially as the number of users increases. This is because, in the broadcast scheme, the group-rate is limited by the worst user in the group and the channel conditions of the worst user will degrade continuously as the number of users increases. However, in the OUS schemes, the rate limitations caused by the worst users are alleviated by selecting users with sufficiently reliable channels in each block. For large K, the subset search algorithm performs better than the proposed scheme in the single-group scenario, but requires significantly higher computational complexity and rapidly becomes intractable as K increases.

Next, we consider the single-group multicasting scenario with $B = 3$ BSs serving collaboratively $K = 30$ users with non-identically distributed channel vectors. We assume that the

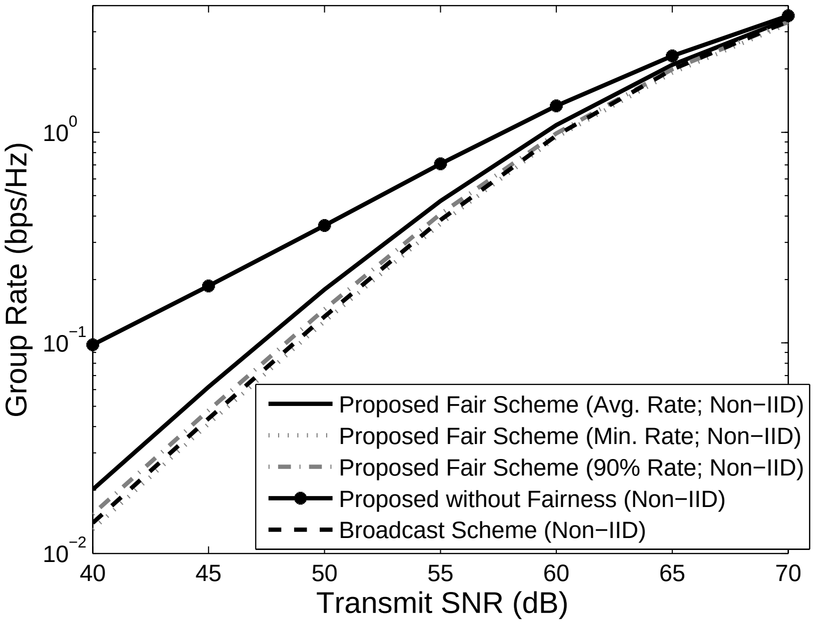

Fig. 7. For the single-group multicasting scenario with $B = 3$ BSs, $M = 2$ antennas per BS, and $K = 30$ non-i.i.d. users, we show the average group-rates of single-group multicasting using the proposed OUS scheme with and without fairness considerations, and the broadcast scheme. The minimum and 90% rates of the proposed fair OUS scheme is also shown for comparison.

users are uniformly distributed in a $[0, 800]$m$\times[0, 733]$m region and the coordinates of the 3 BSs are $(150, 150)$, $(650, 150)$, and $(400, 583)$, respectively. The BS locations are chosen such that they are distanced equally by 500 meters. The channel vector $\mathbf{h}_{b,k}[n]$ is assumed to have entries that are i.i.d. $\mathcal{CN}(0, d_{b,k}^{-\alpha})$ with path loss exponent $\alpha = 2.5$. Here, 6 sets of random user locations are considered, each averaged over 100 channel realizations. In Fig. 7, we show the average group-rates versus the transmit SNR of the proposed single-group multicasting scheme with and without fairness considerations as well as that of the broadcast scheme. The minimum rate among all users (i.e., (7)) as well as the rate achieved by 90% of users (called the 90% rate) are also shown for comparison. The transmit SNR is given by the transmit power P since the noise variance is set as 1. The proposed scheme with fairness considerations refers to the scheme that utilizes normalized channel vectors to compute the user subsets whereas the scheme without fairness refers to the scheme that utilizes the original channel vectors to compute the user subsets. We can see that the proposed fair OUS scheme can still achieve average group-rate that is higher than the broadcast scheme even though a loss is experienced compared to the

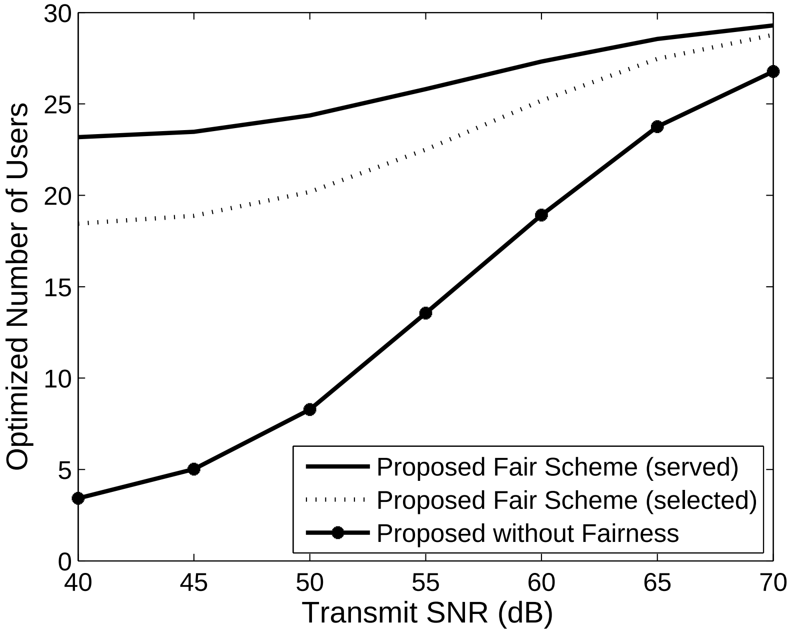

Fig. 8. For the single-group multicasting scenario with $B = 3$ BSs, $M = 2$ antennas per BS, and $K = 30$ non-i.i.d. users, we show the average number of selected and served users obtained with the proposed OUS scheme with and without fairness considerations.

case without fairness. The minimum rate, however, can be lower than the rate achieved by the broadcast scheme if the message is not transmitted over a sufficiently large number of channel realizations. By averaging over 100 channel realization in this figure, the minimum rate is slightly lower than the broadcast scheme due to the diversity of the rates among users. However, at least 90% of users experience rates higher than that of the broadcast scheme, even though our scheme is derived using the average group-rate (instead of the minimum rate in (7)) as the optimization criterion. This demonstrates the effectiveness of the proposed normalization. The minimum rate will become closer to the average rate as the transmission occurs over larger number of channel realizations.

In Fig. 8, we show the average number of users that are selected in the schemes with and without fairness. We can see that, in the case without fairness considerations, less users are selected, and the selected users almost always correspond to users close to the BSs. As mentioned in Section III-B, in the proposed fair OUS scheme, the users that are selected as target users are not the only users that may actually be served. In fact, by setting the rate to satisfy the worst

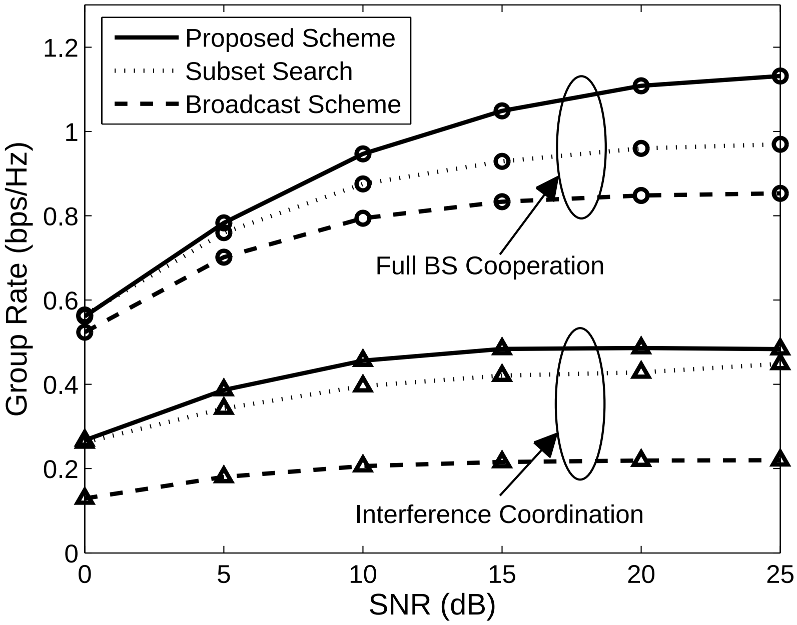

Fig. 9. For the multi-group multicasting scenario with $B = 3$ BSs, $M = 2$ antennas per BS, and $|\mathcal{K}_1| = |\mathcal{K}_2| = |\mathcal{K}_3| = 10$, we show the average group-rates of multi-group multicasting using the proposed, the subset search, the broadcast scheme. The results of both full BS cooperation and interference coordination scenarios are shown.

user in the target user subset, users that are close to the BS (but outside of the target subset) may also have the opportunity to successfully decode. Hence, the number of users served is often greater than the number of users selected. However, we see from Fig. 8 that there is not a significant difference between the two because of the directionality of the signal (i.e., the choice of the transmit covariance matrix). Interestingly, this is different from the single-antenna scenario where the two numbers may have a significant difference since, without spatial directionality, users close to the BS will have a high probability of being served when the target user subset includes a cell-edge user.

B. Multi-Group Multicasting Scenario

In this section, we consider the multi-group multicasting scenario with $B = 3$ BSs serving collaboratively $G = 3$ multicast groups, each with 10 users (i.e., $|\mathcal{K}_1| = |\mathcal{K}_2| = |\mathcal{K}_3| = 10$). In Fig. 9, we show the average group-rate of the relaxed GRB via SCA scheme proposed in Algorithm 4 (c.f. Section IV), and compare it with the subset search algorithm (averaged over only 300 channel realizations) and the broadcast scheme, where all users are served simultaneously in

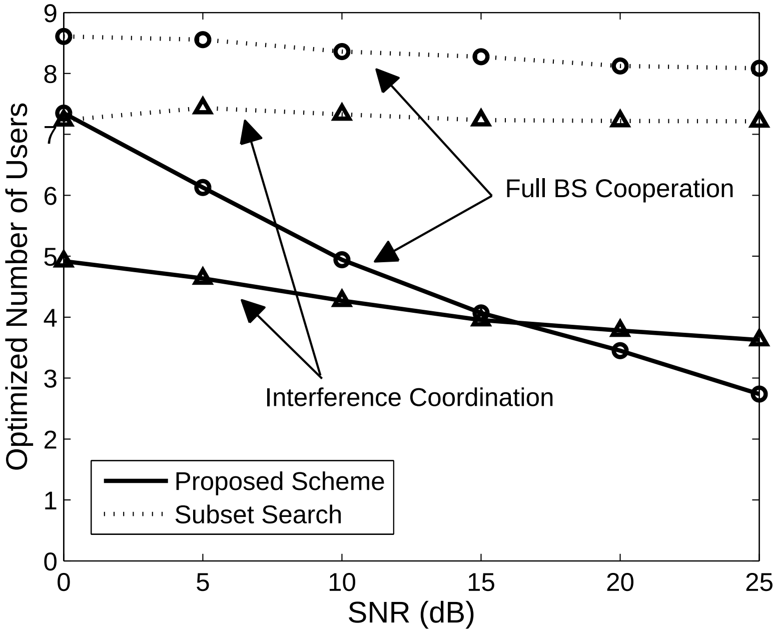

Fig. 10. For the multi-group multicasting scenario with $B = 3$ BSs, $M = 2$ antennas per BS, and $|\mathcal{K}_1| = |\mathcal{K}_2| = |\mathcal{K}_3| = 10$, we show the average number of selected users per group in the proposed and the subset search algorithms. The results of both full BS cooperation and interference coordination scenarios are shown.

each slot. Both cases with full BS cooperation and interference coordination are considered. In the latter case, we assume that each BS serves only one group (namely, BS b serves group g, for $b = g \in \{1, 2, 3\}$) and, thus, $\{\mathbf{Q}_g\}_{b,b'} = \mathbf{0}_{M\times M}$, for all $b \neq b'$ and for all $b \neq g$. Different from the single-group scenario, we can see that user selection in the multi-group scenario is more advantageous in the high SNR regime and the gain is much more significant than that in the single-group case. This is because, in the multi-group scenario, the performance is interference limited at high SNR and, thus, user selection not only can help avoid rate limitations by the user with the worst channel but can also help reduce interference between signals intended for different groups. More interestingly, the proposed scheme also outperforms the subset search algorithm in the multigroup scenario since the latter scheme is more likely to converge towards a locally optimal solution in this scenario. These advantages can be observed in both full BS cooperation and interference coordination scenarios. Interestingly, the group-rate improvement is more significant for the case with only interference coordination. This is because, when BSs are not able to fully cooperate, the spatial degrees of freedom are not sufficient to effectively

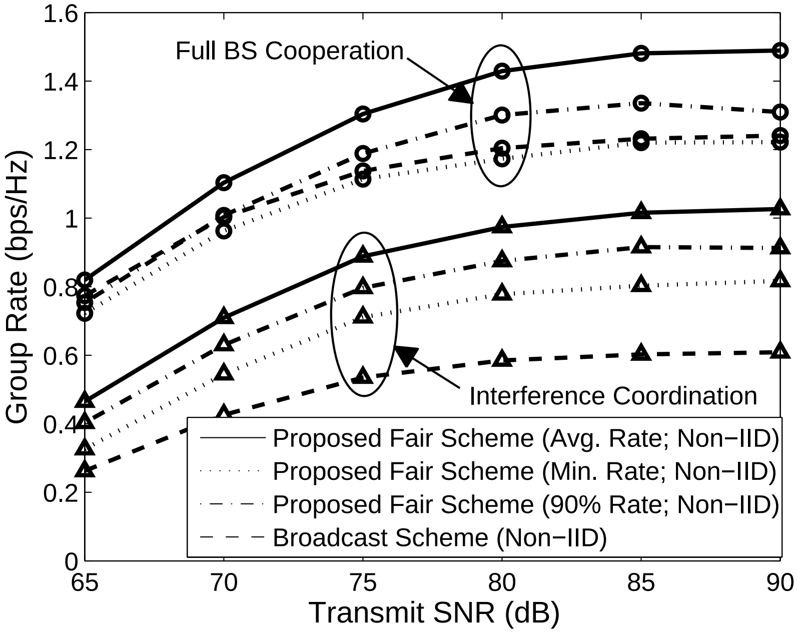

Fig. 11. For the multi-group multicasting scenario with $B = 3$ BSs, $M = 2$ antennas per BS, and $|\mathcal{K}_1| = |\mathcal{K}_2| = |\mathcal{K}_3| = 10$ non-i.i.d. users, we show the average group-rates of multi-group multicasting using the proposed fair OUS scheme and the broadcast scheme. The results of both full BS cooperation and interference coordination scenarios are shown. The minimum and 90% rates of the proposed fair OUS scheme is also shown for comparison.

reduce interference solely through the design of the transmit covariance matrix. Therefore, the benefit of reducing interference through user selection is more pronounced in this case.

In Fig. 10, we show the average number of selected users per group when using the proposed algorithm under full BS cooperation and interference coordination. Interestingly, we can see that, different from the single-group scenario, the number of selected users is less at high SNR instead of at low SNR. This is due to the fact that, at high SNR, the performance is interference limited and, thus, the system would benefit more from eliminating users and reducing interference. This effect is more evident in the case of full BS cooperation where more spatial degrees of freedom are available for signal enhancement and interference avoidance.

In Fig. 11, we consider the multi-group multicasting scenario with users whose channel vectors are non-identically distributed. Again, we have $B = 3$ BSs serving collaboratively $G = 3$ multicast groups, each with 10 users (i.e., $|\mathcal{K}_1| = |\mathcal{K}_2| = |\mathcal{K}_3| = 10$). The users are deployed in the same way as that in Figs. 7 and 8. Again, 6 sets of random user locations are considered, each averaged over 100 channel realizations. Each BS serves a group consisting of the closest 10

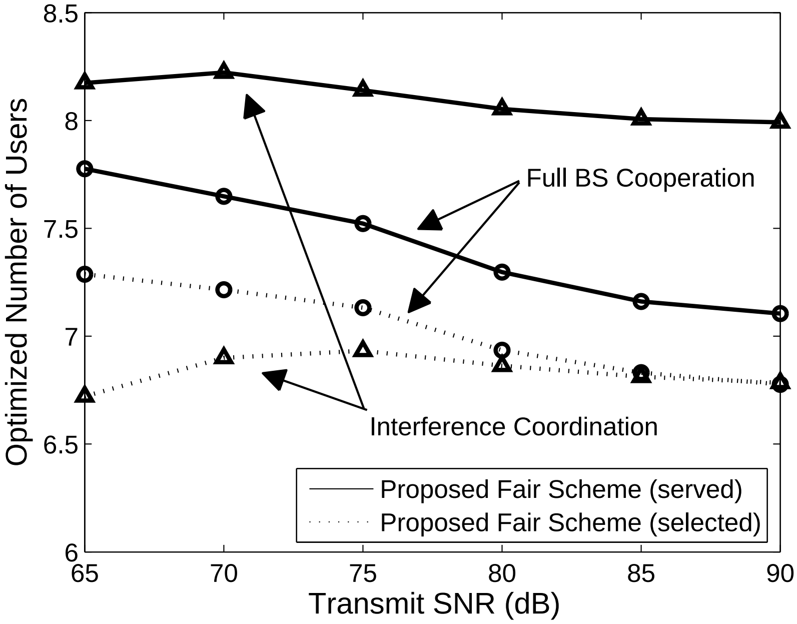

Fig. 12. For the multi-group multicasting scenario with $B = 3$ BSs, $M = 2$ antennas per BS, and $|\mathcal{K}_1| = |\mathcal{K}_2| = |\mathcal{K}_3| = 10$ non-i.i.d. users, we show the average number of selected and served users per group obtained with the proposed fair OUS scheme and the broadcast scheme. The results of both full BS cooperation and interference coordination scenarios are shown.

users. In the figure, we show the average group-rates of the proposed OUS scheme (c.f. Section IV-B) and the broadcast scheme. The minimum rate among all users and the 90% rate are also shown for comparison. Both cases with full BS cooperation and interference coordination are considered. Similar to the i.i.d. case, we can see that the advantages of OUS increase with SNR and the gains are much more significant than the single-group scenario. Moreover, a significant advantage can still be observed in the case of interference coordination. This is because the normalized channel vectors used in the proposed OUS scheme preserves the direction of the channel vectors and, thus, is still able to successfully perform interference coordination among different BSs. Similar to the single-group scenario, the minimum rate may be smaller than the rate achieved in the broadcast scheme if the transmission does not occur over a large number of time slots, which is the case in the full BS cooperation scenario. However, the majority of users (in fact, over 90% of users) achieve rates that are higher than that of the broadcast scheme. In Fig. 12, we show the average number of selected and served users obtained using the proposed fair OUS scheme under both full BS cooperation and interference coordination. We can see that the number of served users is again more than that of selected target users. However, this effect

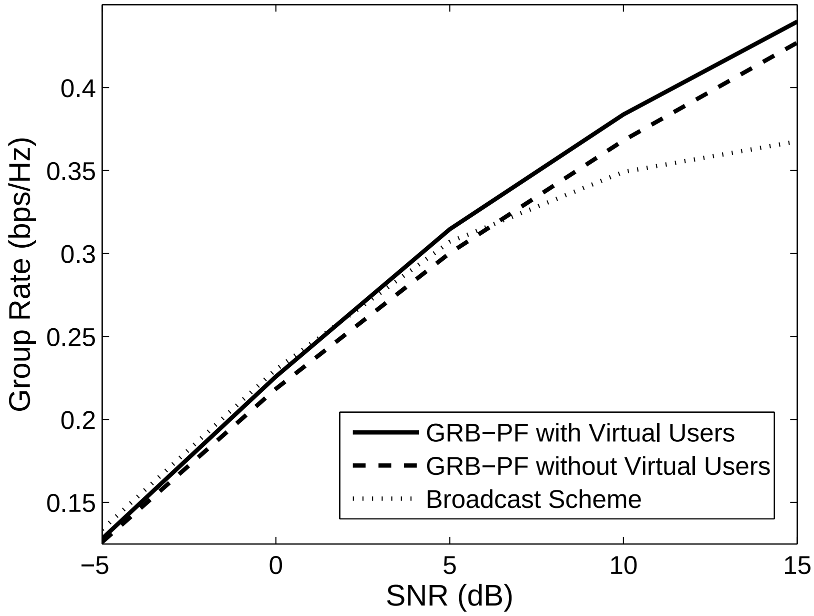

Fig. 13. Multi-group multicasting scenario with $B = 3$ BSs, $M = 2$ antennas per BS, and $|\mathcal{K}_1| = |\mathcal{K}_2| = |\mathcal{K}_3| = 20$, among which only 5 per group feedback their channel vectors. We plot the average group-rate of the cases with and without virtual users.

is less pronounced under full BS cooperation since the signal in this case contains directionality.

Finally, let us consider the GRB-PF problem as discussed in Section V. Here, we assume that there are 20 users per group (i.e., $|\mathcal{K}_1| = |\mathcal{K}_2| = |\mathcal{K}_3| = 20$), but only 5 per group feedback their channel vectors in each block. Notice that, in addition to the proposed scheme, it is also possible to compute the input covariance matrix, the rate, and the user selection assuming that only the users who feedback their CSI exist in the network. The latter is referred to as the case without virtual users. In the experiments, the proposed scheme is implemented with $J = 100$ virtual users. In Fig. 13, we show the average group-rate achieved when the system parameters are derived using the above two approaches. We can see that, by considering virtual users, the group-rate can be significantly improved, especially at high SNR where user selection is critical. The performance of the broadcast scheme where the system parameters are designed by serving all feedback users in each block is also plotted for comparison. We can see that, at high SNR, where the performance is interference limited, user selection (even without consideration of virtual users) can provide significant group-rate improvement.

VII. Conclusion

In this work, the OUS scheme was examined for both single-group and multi-group multicasting scenarios in the physical layer of a multicell multi-antenna wireless system. User selection along with application layer erasure coding was proposed to overcome rate limitations caused by the worst user in the multicast group. The proposed user selection policies were derived based on the optimization and relaxation of a set of user selection variables. For the single group scenario, we formulated the problem as a group-rate maximization problem, and proposed an efficient user selection policy by performing a convex relaxation and by employing a sequential deflation algorithm. For the multi-group scenario, we formulated the problem as a group-rate balancing problem and proposed an efficient user selection policy by performing SCA along with the sequential deflation algorithm. When only part of the users feedback their instantaneous CSI, we further introduced the concept of virtual users to take into consideration the probability that non-feedback users are served in each block. The effectiveness of the proposed schemes was shown via computer simulations. Interestingly, we observed that user selection is more advantageous in the low SNR regime for the single-group scenario, but is more beneficial in the high SNR regime for the multi-group scenario.

Appendix A
Proof of Lemma 1

We first show that, for $\delta \leq \left[\max_k \log_2(1+\sum_{b=1}^{B} P_b \|\mathbf{h}_{b,k}\|^2)\right]^{-1}$, $(\mathbf{Q}, R, \{s_k\}_{k\in\mathcal{K}})$ is a feasible point of (13) if and only if $(\mathbf{Q}, R, \mathcal{A})$, where $\mathcal{A} = \{k \in \mathcal{K} : s_k = 1\}$, is a feasible point of (10). Specifically, let $(\mathbf{Q}, R, \{s_k\}_{k\in\mathcal{K}})$ be a feasible point of (13) and let $\mathcal{A} = \{k \in \mathcal{K} : s_k = 1\}$. We can see that, for any $k \in \mathcal{A}$, the constraint in (10b) is equivalent to the constraint in (13b) when $s_k = 1$. Therefore, if $(\mathbf{Q}, R, \{s_k\}_{k\in\mathcal{K}})$ is a feasible point of (13), then $(\mathbf{Q}, R, \mathcal{A})$ must be a feasible point of (10b) as well. On the other hand, let $(\mathbf{Q}, R, \mathcal{A})$ be a feasible point of (10b) and let $s_k = 1$ if $k \in \mathcal{A}$, and $s_k = 0$, if $k \notin \mathcal{A}$. Similarly, the constraints in (13b) are the same as those in (10b), for $k \in \mathcal{A}$ (i.e., for k such that $s_k = 1$). However, for $k' \notin \mathcal{A}$ (i.e., for k' such that $s_{k'} = 0$), the constraints in (13b) are redundant since $\log_2[1 + \text{tr}(\mathbf{Q}\mathbf{h}_{k'}\mathbf{h}_{k'}^H)] + \delta^{-1}(1 - s_{k'}) \geq \log_2[1 + \text{tr}(\mathbf{Q}\mathbf{h}_k\mathbf{h}_k^H)] \geq R$, for all $k \in \mathcal{A}$. Therefore, $(\mathbf{Q}, R, \{s_k\}_{k\in\mathcal{K}})$ is a feasible point of (13) if $(\mathbf{Q}, R, \mathcal{A})$ is a feasible point of (10b). Moreover, one can also see that

$(\mathbf{Q}, R, \mathcal{A})$ and $(\mathbf{Q}, R, \{s_k\}_{k\in\mathcal{K}})$ achieve the same objective values in their respective problems since $\frac{1}{|\mathcal{K}|}R\sum_{k\in\mathcal{K}} s_k = \frac{1}{|\mathcal{K}|}R\sum_{k\in\mathcal{K}} \mathbf{1}_{\{k\in\mathcal{A}\}}$. The lemma follows.

Appendix B
Proof of Proposition 1

We basically show that Algorithm 4 is a special case of the successive upper-bound minimization (SUM) method in [39].

First, notice that problem (21) is the epigraph form of the following problem:

$$\max\ \min_g \left\{ \frac{\sum_{k\in\mathcal{K}_g} s_k}{\tau_g|\mathcal{K}_g|} \min_{k\in\mathcal{K}_g} \left[r_k(\{\mathbf{Q}_\ell\}_{\ell=1}^G) + \delta^{-1}(1-s_k) \right] \right\} \tag{36a}$$

$$\text{subject to}\ \ 0\le s_k \le 1,\ \forall k\in\mathcal{K},\ \text{(20c)},\ \text{and (20d)}, \tag{36b}$$

$$\text{variables: } \{\mathbf{Q}_g\}_{g=1}^G, \{s_k\}_{k\in\mathcal{K}}. \tag{36c}$$

Since R_g and α are auxiliary variables, we can focus on showing that any limit point of $\{\{\mathbf{Q}_g^*\}_{g=1}^G, \{s_k^*\}_{k=1}^K\}$ is a stationary point of problem (36). Similar to problem (21), problem (25) is the epigraph form of

$$\max\ \min_g \left\{ \frac{\sum_{k\in\mathcal{K}_g} s_k}{\tau_g|\mathcal{K}_g|} \min_{k\in\mathcal{K}_g} \left[\bar{r}_k(\{\mathbf{Q}_\ell\}_{\ell=1}^G \,|\, \{\tilde{\mathbf{Q}}_\ell\}_{\ell=1}^G) + \delta^{-1}(1-s_k) \right] \right\} \tag{37a}$$

$$\text{subject to}\ \ 0\le s_k \le 1,\ \forall k\in\mathcal{K},\ \text{(20c)},\ \text{and (20d)}, \tag{37b}$$

$$\text{variables: } \{\mathbf{Q}_g\}_{g=1}^G, \{s_k\}_{k\in\mathcal{K}}. \tag{37c}$$

The approximation $r_k(\{\mathbf{Q}_\ell\}_{\ell=1}^G) \approx \bar{r}_k(\{\mathbf{Q}_\ell\}_{\ell=1}^G \mid \{\tilde{\mathbf{Q}}_\ell\}_{\ell=1}^G)$ satisfies $r_k(\{\tilde{\mathbf{Q}}_\ell\}_{\ell=1}^G) = \bar{r}_k(\{\tilde{\mathbf{Q}}_\ell\}_{\ell=1}^G \mid \{\tilde{\mathbf{Q}}_\ell\}_{\ell=1}^G)$, $r_k(\{\mathbf{Q}_\ell\}_{\ell=1}^G) \ge \bar{r}_k(\{\mathbf{Q}_\ell\}_{\ell=1}^G \mid \{\tilde{\mathbf{Q}}_\ell\}_{\ell=1}^G)$, and

$$\left.\frac{\partial r_k(\{\mathbf{Q}_\ell\}_{\ell=1}^G)}{\partial \mathbf{Q}_g}\right|_{\mathbf{Q}_\ell=\tilde{\mathbf{Q}}_\ell,\forall\ell} = \left.\frac{\partial r_k(\{\mathbf{Q}_\ell\}_{\ell=1}^G \mid \{\tilde{\mathbf{Q}}_\ell\}_{\ell=1}^G)}{\partial \mathbf{Q}_g}\right|_{\mathbf{Q}_\ell=\tilde{\mathbf{Q}}_\ell,\forall\ell},$$

for all g. Therefore, Algorithm 4 is essentially the SUM method [39]. According to [39, Theorem 1], any limit point of $\{\{\mathbf{Q}_g^*\}_{g=1}^G, \{s_k^*\}_{k=1}^K\}$ generated by Algorithm 4 is a stationary point of (36). Proposition 1 is thus proved.

REFERENCES

[1] M. Gruber and D. Zeller, “Multimedia broadcast multicast service: New transmission schemes and related challenges,” *IEEE Commun. Mag.*, vol. 49, no. 12, pp. 176–181, Dec. 2011.

[2] D. Lecompte and F. Gabin, “Evolved multimedia broadcast/multicast service (eMBMS) in LTE-advanced: Overview and Rel-11 enhancements,” *IEEE Commun. Mag.*, vol. 50, no. 11, pp. 68–74, Nov. 2012.

[3] T. Jiang, W. Xiang, H.-H. Chen, and Q. Ni, “Multicast broadcast services support in OFDMA-based WiMAX systems,” *IEEE Commun. Mag.*, vol. 45, no. 8, pp. 78–86, Aug. 2007.

[4] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, “Transmit beamforming for physical-layer multicasting,” *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, June 2006.

[5] A. Lozano, “Long-term transmit beamforming for wireless multicasting,” in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 3, 2007, pp. III–417–III–420.

[6] H. Zhu, N. Prasad, and S. Rangarajan, “Precoder design for physical layer multicasting,” *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5932–5947, Nov. 2012.

[7] N. Jindal and Z.-Q. Luo, “Capacity limits of multiple antenna multicast,” in *Proc. IEEE Intl. Symp. Inform. Theory (ISIT)*, 2006, pp. 1841–1845.

[8] E. Karipidis, N. D. Sidiropoulos, and Z.-Q. Luo, “Quality of service and max-min fair transmit beamforming to multiple cochannel multicast groups,” *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1268–1279, Mar. 2008.

[9] Y. C. B. Silva and A. Klein, “Linear transmit beamforming techniques for the multigroup multicast scenario,” *IEEE Trans. Veh. Technol.*, vol. 58, no. 8, pp. 4353–4367, Oct. 2009.

[10] N. Bornhorst and M. Pesavento, “An iterative convex approximation approach for transmit beamforming in multi-group multicasting,” in *Proc. IEEE Intl. Workshop on Signal Process. Adv. in Wireless Commun. (SPAWC)*, June 2011, pp. 426–430.

[11] D. Christopoulos, S. Chatzinotas, and B. Ottersten, “Weighted fairness multicast multigroup beamforming under per-antenna power constraints,” *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 5132–5142, Oct. 2014.

[12] D. Christopoulos, S. Chatzinotas, and B. Ottersten, “Sum rate maximizing multigroup multicast beamforming under per-antenna power constraints,” in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, Austin, TX, USA, Dec. 2014, preprint: arXiv:1407.0005 [cs.IT].

[13] Z. Xiang, M. Tao, and X. Wang, “Coordinated multicast beamforming in multicell networks,” *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 12–21, Jan. 2013.

[14] G. Dartmann, X. Gong, and G. Ascheid, “Low complexity cooperative multicast beamforming in multiuser multicell downlink networks,” in *Proc. Intl. ICST Conf. on Cognitive Radio Oriented Wireless Networks and Commun. (CROWNCOM)*, 2011, pp. 370–374.

[15] U. C. Kozat, “On the throughput capacity of opportunistic multicasting with erasure codes,” in *Proc. IEEE Intl. Conf. on Computer Commun. (INFOCOM)*, Apr. 2008, pp. 520–528.

[16] P. K. Gopala and H. E. Gamal, “Opportunistic multicasting,” in *Proc. Asilomar Conf. on Signals, Systems and Computers*, Nov. 2004, pp. 845–849.

[17] ——, “On the throughput-delay tradeoff in cellular multicast,” in *Proc. IEEE Intl. Conf. on Wireless Networks, Commun. and Mobile Computing*, vol. 2, June 2005, pp. 1401–1406.

[18] T.-P. Low, M.-O. Pun, Y.-W. P. Hong, and C.-C. J. Kuo, “Optimized opportunistic multicast scheduling (OMS) over wireless cellular networks,” *IEEE Trans. Wireless Commun.*, vol. 9, no. 2, pp. 791–801, Feb. 2010.

[19] M. Luby, T. Gasiba, T. Stockhammer, and M. Watson, “Reliable multimedia download delivery in cellular broadcast networks,” *IEEE Trans. Broadcast.*, vol. 53, no. 1, pp. 235–246, Mar. 2007.

[20] D. J. C. MacKay, “Fountain codes,” *IEE Proceedings-Communications*, vol. 152, no. 6, pp. 1062–1068, Dec. 2005.

[21] T.-P. Low, P.-C. Fang, Y.-W. P. Hong, and C.-C. J. Kuo, “Multi-Antenna Multicasting with Opportunistic Multicast Scheduling and Space-Time Transmission,” in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2010.

[22] M. Kaliszan, E. Pollakis, and S. Stanczak, “Efficient beamforming algorithms for MIMO multicast with application-layer coding,” in *Proc. IEEE Intl. Symp. on Inform. Theory (ISIT)*, 2011, pp. 928–932.

[23] ——, “Multigroup multicast with application-layer coding: Beamforming for maximum weighted sum rate,” in *Proc. IEEE Wireless Commun. and Networking Conference (WCNC)*, 2012, pp. 2270–2275.

[24] E. Matskani, N. D. Sidiropoulos, Z.-Q. Luo, and L. Tassiulas, “Convex approximation techniques for joint multiuser downlink beamforming and admission control,” *IEEE Trans. Wireless Commun.*, vol. 7, no. 7, pp. 2682–2693, July 2008.

[25] ——, “Efficient batch and adaptive approximation algorithms for joint multicast beamforming and admission control,” *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 4882–4894, Dec. 2009.

[26] W. Wang and S. Ahmed, “Sample average approximation of expected value constrained stochastic programs,” *Operations Research Letters*, vol. 36, no. 5, pp. 515 – 519, 2008.

[27] B. K. Pagnoncelli, S. Ahmed, and A. Shapiro, “Sample average approximation method for chance constrained programming: Theory and applications,” *Journal of Optimization Theory and Applications*, vol. 142, no. 2, pp. 399–416, 2009.

[28] D. Gesbert, S. Hanly, H. Huang, S. S. Shitz, O. Simeone, and W. Yu, “Multi-cell MIMO cooperative networks: A new look at interference,” *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010.

[29] E. Björnson, R. Zakhour, D. Gesbert, and B. Ottersten, “Cooperative multicell precoding: Rate region characterization and distributed strategies with instantaneous and statistical CSI,” *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4298–4310, Aug. 2010.

[30] H. Dahrouj and W. Yu, “Coordinated beamforming for the multicell multi-antenna wireless system,” *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, pp. 1748–1759, May 2010.

[31] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, “Semidefinite relaxation of quadratic optimization problems,” *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.

[32] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley-Interscience, 2006.

[33] M. Schubert and H. Boche, “Solution of the multiuser downlink beamforming problem with individual SINR constraints,” *IEEE Trans. Veh. Technol.*, vol. 53, no. 1, pp. 18–28, Jan. 2004.

[34] M. Grant and S. Boyd, “CVX: Matlab software for disciplined convex programming, version 2.0 beta,” http://cvxr.com/cvx, Sept. 2012.

[35] S. P. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[36] W.-C. Li, T.-H. Chang, C. Lin, and C.-Y. Chi, “Coordinated beamforming for multiuser MISO interference channel under rate outage constraints,” *IEEE Trans. Signal Process.*, vol. 61, no. 5, pp. 1087–1103, Mar. 2013.

[37] S.-M. Huang, J.-N. Hwang, and Y.-C. Chen, “Reducing feedback load of opportunistic multicast scheduling over wireless systems,” *IEEE Commun. Lett.*, vol. 14, no. 12, pp. 1179–1181, Dec. 2010.

[38] M. Li, X. Wang, D. Wang, and J. Zhou, “Feedback load reduction scheme in OFDM-based wireless multicast systems,” in *Proc. IEEE Wireless Commun. and Networking Conf. (WCNC)*, 2013, pp.1068,1072.

[39] M. Razaviyayn, M. Hong, and Z.-Q. Luo, “A unified convergence analysis of block successive minimization methods for nonsmooth optimization,” *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 1126–1153, 2013.